



UNIVERSITÉ DE NANTES



**IAE NANTES**  
ÉCONOMIE & MANAGEMENT

ANALYSE DES VARIABLES QUALITATIVES

---

## **Analyse des pratiques culturelles des étudiants nantais - deuxième volet**

---



Laurianne MORICEAU  
Diane THIERRY

Enseignante : Mme. TRAVERS  
Année universitaire : 2020-2021

Master 2 Économétrie et Statistiques, parcours Économétrie Appliquée

## Avant-propos

Ce dossier est la continuité d'une analyse réalisée en M1 EKAP par Diane Thierry et Laurianne Moriceau en 2020, dans le cadre du cours de Variables Qualitatives 1. La base de données sur laquelle nous travaillerons est ainsi la même que l'année dernière, à savoir les réponses que nous avons récoltées d'un questionnaire diffusé aux étudiants de Nantes pour connaître leurs pratiques culturelles. L'objectif cette fois-ci est de prendre en considération les remarques faites sur le premier compte rendu, afin de proposer une analyse plus approfondie. Nous avons également prêté attention à ne pas nous répéter en proposant de nouveaux angles d'études dans l'introduction tout comme dans la partie économique, pour que ce dossier ne soit pas rébarbatif compte tenu du fait que le sujet a déjà été présenté.

## Résumé

Après une première analyse sur les déterminants des pratiques culturelles des étudiants de Nantes où nous cherchions à expliquer la probabilité de se rendre au théâtre d'une part, et les fréquences des pratiques culturelles d'autre part, nous avons décidé de nous replonger dans ce sujet en complétant notre étude. En outre, nous nous intéressons ici à modéliser conjointement plusieurs phénomènes de manière à mettre en évidence les liens et les influences qui peuvent exister entre les variables à expliquer. Dans un premier temps nous avons cherché à modéliser un modèle biprobit expliquant la fréquentation du théâtre conjointement à la fréquentation des conférences. Cependant, le paramètre qui matérialise le lien entre les 2 probits noté  $\rho$  n'étant pas significatif, nous avons réalisé 2 logits séparés. Ensuite, nous avons cherché à modéliser l'existence supposée d'une relation linéaire du budget sur la fréquentation du théâtre et de l'influence de l'âge sur les conférences. À partir des modèles logit nous avons appliqué des modèles linéaires généralisés. Les résultats montrent qu'il existe bien des effets de seuils du budget sur la probabilité de fréquentation du théâtre. Ainsi, dans le cas du théâtre, le modèle *GAM* permet effectivement d'améliorer les prévisions. En revanche l'amélioration n'est pas significative concernant les conférences, puisqu'il n'existe pas de non-linéarité avec l'âge. Pour terminer, nous avons modélisé un triprobit sur la pratique d'activités quotidiennes ; la lecture, les jeux vidéo ainsi que la radio. Le paramètre  $\rho$  était cette fois significatif et nous avons pu conclure sur l'existence d'un lien entre ces phénomènes. Les résultats montrent que la probabilité de lire au moins une fois par semaine dépend positivement de l'intérêt pour les bibliothèques, la musique classique et la danse au seuil de 1%. La probabilité d'écouter la radio dépend elle de l'intérêt pour le théâtre et à la participation des étudiants aux activités culturelles universitaires au seuil de 5%. Enfin, la probabilité de jouer au moins une fois par semaine aux jeux vidéo dépend des caractéristiques de l'individu, dont le genre et l'âge.

# Abstract

After a first analysis on the determinants of the cultural practices of the students of Nantes where we sought to explain going or not to the theater on the one hand, and the frequencies of cultural practices on the other hand, we decided to dive back into this subject by completing our study. Thus, we are interested here to jointly model several phenomena, so as to highlight the links and influences that may exist between the variables to be explained (regressands). Initially, we sought to model a biprobit model which explains theater attendance together with conference attendance. However, given that the parameter which materializes the link between the 2 probits noted  $\rho$  wasn't significant, we carried out 2 separate logits. Then, in order to explain the non-linear relationship of 'budget' on theater attendance, and 'age' on lectures assumed from logit models, we applied generalized additive models. In the case of the theater, the GAM effectively improves forecasts, on the other hand the improvement was not significant for conferences. To finish, we modeled a triprobit on the practice of daily activities such as reading, listening radio or playing video games. The  $\rho$  parameter was this time significant so we were able to conclude on the existence of a link between these phenomena. It then appeared that the probability of reading at least once a week is positively dependent on interest in libraries, classical music and dance at the 1% threshold. The probability of listening to the radio depends on the interest in the theater and on the participation of the students in university cultural activities at the threshold of 5%. Finally, the probability of playing video games at least once a week depends on the characteristics of the individual, such as gender and age.

## Sommaire

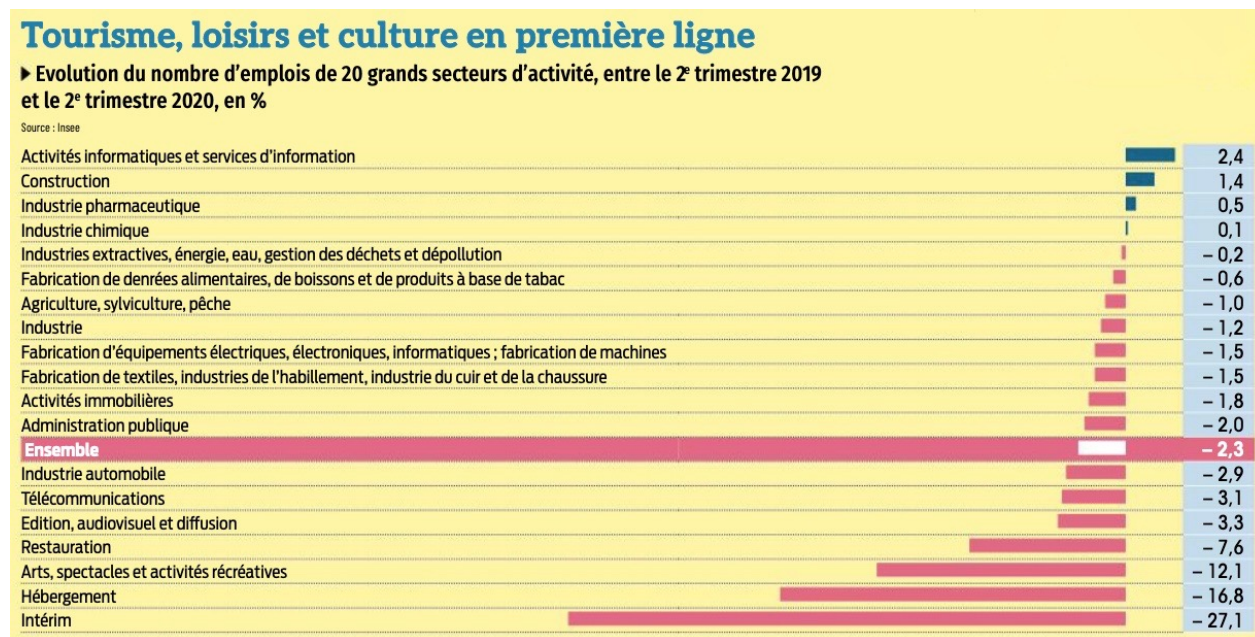
<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>L'enquête</b>	<b>7</b>
<b>3</b>	<b>Analyse préliminaire</b>	<b>12</b>
<b>4</b>	<b>Modèle biprobit</b>	<b>20</b>
<b>5</b>	<b>Modèle additif généralisé</b>	<b>25</b>
<b>6</b>	<b>Modèle probit multivarié</b>	<b>36</b>
<b>7</b>	<b>Conclusion</b>	<b>40</b>
<b>8</b>	<b>Annexes</b>	<b>42</b>
<b>9</b>	<b>Références bibliographiques</b>	<b>48</b>

# 1 Introduction

L'année précédente, nous lançons une enquête auprès des étudiants nantais afin de rendre compte de leurs pratiques culturelles. Bien que différentes dans ses formes et ses accès, les pratiques culturelles sont pour nous essentielles au bon équilibre d'une société puisqu'elles lient les individus par un sentiment de cohésion et d'appartenance. Elle permet également une ouverture d'esprit au niveau individuel. La culture est parfois définie ou interprétée en sociologie comme la séparation avec la nature ; elle est alors synonyme d'éducation de l'esprit et de civilisation, et par conséquent d'Humanisme.

La France, comme la plupart des pays du monde, connaît une période difficile liée à la crise initialement sanitaire qui s'est ensuite transformée en crise économique, puis en crise sociale touchant majoritairement les moins de 25 ans<sup>1</sup>. Durant cette crise et les périodes de confinement, les commerces et activités jugés non indispensables à la vie du pays ont été contraints de fermer. Ainsi, en cette période de repli social, le secteur culturel dont les arts, les spectacles et les activités récréatives, est l'un de ceux qui a été le plus éprouvé par ces restrictions. On voit en effet, sur le graphique suivant, que le nombre d'emplois dans le secteur culturel a diminué de 12,1% au 2<sup>e</sup> trimestre 2020 par rapport à l'année précédente. Pour l'ensemble des secteurs, ce taux est près de 4 fois plus faible puisqu'en moyenne les emplois au 2<sup>e</sup> trimestre 2019 ont diminué de 2,3% par rapport au même trimestre de l'année 2019.

FIGURE 1 – Evolution du nombre d'emplois de 20 grands secteurs d'activité



La ministre de la Culture Roselyne Bachelot a ainsi annoncé, le 22 octobre dernier lorsque le couvre-feu s'était imposé dans plusieurs grandes villes de France, une aide spéciale pour les acteurs du secteur du spectacle vivant et du cinéma. Ces sommes s'élèvent respectivement à 85 millions et 30 millions d'euros - affirmant que "Chacun a besoin, nous avons besoin de culture, et peut être encore plus durant cette crise qui a affecté notre capacité à nous rassembler".<sup>2</sup>

1. MENS, "Crise : qui portera quel chapeau ?"

2. CLARISSE et SANDRINE, *Couvre-feu : 115 millions d'euros pour soutenir le spectacle vivant et le cinéma*.

Les pratiques culturelles sont-elles importantes pour la santé d'un pays ? Bien qu'en mauvaise posture depuis la crise mondiale, nous souhaitons approfondir davantage notre analyse commencée l'année dernière en cernant un plus grand nombre de pratiques culturelles. De fait, nous avons étudié les déterminants des sorties au théâtre, et nos principales conclusions résidaient en ce que les étudiants sensibilisés et habitués aux sorties culturelles avaient une probabilité plus élevée de s'y rendre. En revanche, regarder la télévision diminuait cette probabilité. De la même manière, la fréquence d'activité culturelle augmentait avec le budget et le niveau d'étude de l'individu.

Nos résultats semblaient satisfaisants, mais nous souhaitons mettre en évidence des liens entre les pratiques grâce aux nouveaux outils statistiques étudiés cette année, tels que le biprobit, le triprobit ou encore le modèle additif généralisé. Notre objectif est ainsi de montrer que les pratiques culturelles sont généralement le résultat d'habitudes liées à la sensibilisation et la familiarisation et qu'elles sont liées entre elles - en expliquant par différentes modélisations la fréquence d'une activité par rapport à la fréquence d'autres activités.

Ainsi, nous verrons dans une première partie la présentation des variables issues de notre questionnaire en distinguant les variables à expliquer dans un premier temps, puis les déterminants qui peuvent rendre compte de la pratique ou de la non pratique des activités culturelles. Ensuite, nous reprendrons une analyse descriptive de notre base afin de mieux appréhender les modèles de régression, par la mise en relation, par exemple, des variables que nous chercherons à expliquer avec certains facteurs qui nous semblent incontournables. Nous commencerons alors l'analyse économétrique en modélisant à l'aide d'un probit bivarié le fait de se rendre au théâtre conjointement au fait de se rendre à des conférences puis, étant non liés, nous estimerons ces phénomènes à l'aide de logits distincts. Nous chercherons alors à améliorer la qualité de la prévision de nos  $Y_i$  en construisant des modèles additifs généralisés. Enfin, nous terminerons cette étude en appliquant un modèle probit trivarié sur les fréquences de pratiques culturelles du quotidien (que nous distinguerons en première partie des **sorties** culturelles), que sont la lecture, la radio et les jeux vidéo.

## 2 L'enquête

La culture étant pour nous une porte d'ouverture sur le monde, un moyen d'enrichir ses savoirs et de se construire, nous souhaitons savoir quelles étaient ces pratiques chez les étudiants. Nous avons donc recueilli de nombreuses informations, notamment les types d'activités culturelles pratiquées, les fréquences, les volontés de changer ou diversifier ces pratiques, et enfin les caractéristiques propres à chaque étudiant. Nous avons obtenu un total de 350 observations et 33 variables explicatives. Cette enquête réalisée du 29 janvier au 10 février 2020 nous a permis d'obtenir un large choix d'informations que nous avons traitées afin de mettre en évidence les tenants et aboutissants des pratiques culturelles. L'année dernière, nous avons mis en pratique la régression logistique, ainsi qu'un modèle multinomial ordonné. L'objectif de ce dossier est d'exploiter à nouveau cette base riche en informations pour pousser davantage l'analyse des pratiques culturelles.

Nous commencerons dans cette partie par expliciter les variables que nous chercherons à expliquer, puis nous nous intéresserons aux déterminants extrasèques des pratiques culturelles en regardant dans un premier temps les déterminants liés aux habitudes culturelles. Puis, nous nous concentrerons sur les déterminants intrasèques liés aux individus et leurs caractéristiques, qui sont eux, indépendants du sujet traité.

### 2.1 Expliquer la fréquence des activités

La première partie de notre questionnaire consistait à demander aux étudiants la fréquence à laquelle ils se rendaient dans les diverses institutions culturelles et ainsi leurs habitudes de **sorties**. Ces sorties induisent un coût d'entrée et se pratiquent principalement le week-end, telles que l'opéra, les musées, les concerts. Dans un second temps, nous leur avons demandé quelles étaient leurs habitudes culturelles pratiquées **depuis chez eux**, telles que la fréquence de visionnage de la télévision, d'écoute de la radio ou de lecture par exemple. Allant de "jamais", à "au moins 1 fois par semaine", les fréquences nous aident ainsi à visualiser les pratiques les plus courantes, et différencier les pratiques dites occasionnelles des quotidiennes.

Plus accessible et pratiquée quotidiennement, cette deuxième catégorie d'activité fait partie intégrante de notre culture. Même si elles ne cochent pas les cases d'une culture plus traditionnelle, ces activités du quotidien n'induisent pas de coût supplémentaire d'entrée comme pour les activités culturelles. Il n'y a ainsi pas ou peu de biais lié à l'accessibilité, et les fréquences sont réellement représentatives des envies. En effet, en France en 2019 88% de la population était connectée à internet<sup>3</sup>, ce chiffre connaît une hausse quasi constante depuis les années 2000 marquées par l'explosion de la bulle internet. De même, selon l'INSEE près de 96% de la population de 16 ans et plus est équipée d'un téléviseur dans son ménage, en revanche ce taux a tendance à diminuer chez les jeunes ; il était de 79% en 2018 contre 97% en 2010.<sup>4</sup> En moyenne le coût d'un tel équipement varie entre 20 et 30€ par mois pour internet et la télévision<sup>5</sup> - auquel il faut ajouter le coût du téléviseur à l'achat qui se situe autour de 100€ pour les plus basics.

Il n'y a ainsi pas ou très peu d'inégalités d'accès aux pratiques culturelles du quotidien car la plupart des ménages sont équipés d'internet et d'un téléviseur - qui sont aujourd'hui considérés comme des équipements de base que l'on trouve dans tous types d'habitations, au même titre que les meubles, le four etc (bien que ce chiffre soit moins important chez les jeunes, comme nous avons pu le voir précédemment).

---

3. JULIE, PATRICIA et VICTOR, "Baromètre du numérique 2019".

4. INSEE, *Tableaux de l'économie française - Équipement des ménages*.

5. QUECHOISIR, *Comparateur des Fournisseurs d'accès à Internet*.



Par ailleurs, 64% des jeunes de 18 à 24 ans déclarent sortir au moins une fois par mois, taux alors largement supérieur à celui des 35-49 ans ou des 50-64 ans qui s'élève à seulement 39%.<sup>6</sup> Le coût de telles sorties est un facteur non négligeable car même si les étudiants disposent de réductions sur certains événements culturels, celles-ci ne sont pas toujours connues et ne suffisent parfois pas à favoriser les sorties. Ainsi, l'enquête réalisée par l'institut LH2 révèle que les 15-24 ans - bien que les plus "gourmands" en termes de pratiques culturelles, sont ceux qui dépensent le moins d'argent pour cela ; leur budget alloué aux sorties est à 80% inférieur à 50€ par mois. Une variable explicative faisant référence aux dépenses de chaque étudiant pour les activités culturelles est donc primordiale.

Nous avons ainsi montré que les activités culturelles à la maison sont distinguables des sorties culturelles, car elles ne se pratiquent pas à la même fréquence et n'impliquent pas les mêmes coûts. Seulement, afin d'appliquer nos modèles, les variables à expliquer doivent être équiréparties entre la pratique de l'activité et la non-pratique. Par exemple, le fait de regarder la télévision est une activité largement pratiquée par la plupart des observations de notre échantillon. Ainsi, nous nous concentrerons dans un premier temps sur la fréquentation du théâtre et la fréquentation de conférences ; activités pour lesquelles les fréquences sont bien réparties entre la pratique (régulière comme occasionnelle) et la non-pratique.

## 2.2 Les déterminants des fréquences de pratiques culturelles

Nous distinguons dans les déterminants, les variables qui sont liées à la pratique d'activités culturelles et celles liées à l'identité du répondant et donc indépendantes du sujet traité. Nous verrons les variables explicatives de ces deux catégories successivement.

### 2.2.1 L'offre, l'accessibilité et les habitudes

#### 1. L'offre

La première question que nous nous sommes posée sur le fait de pratiquer ou non une activité culturelle est l'offre qui est proposée. En effet, nous savons que pour se rendre au théâtre, ou à des concerts par exemple, il faut soit chercher l'information sur les activités qui ont lieu, soit que l'information vienne spontanément à nous. Ainsi, nous avons d'abord cherché à savoir si les répondants participent aux sorties proposées par leur établissement. Effectivement, dans la plupart des écoles ou universités se trouvent un bureau des activités (BDA) ou un pôle proposant aux étudiants de participer à certains événements autour de la culture : par exemple les Journées Arts et Culture dans l'Enseignement Supérieur (JACES) qui se déroulent chaque année en France depuis 2013. Dans ce cas-ci, nous sommes certains que l'individu a été confronté à l'information sur les événements culturels à un moment ou un autre - que ce soit par des mails d'invitation, le bouche-à-oreille, les affiches sur le campus... En outre, sa participation ou non résulte du seul fait de sa volonté, tout en sachant que l'offre proposée par ces associations sont adaptées aux goûts et habitudes des étudiants puisque c'est la seule population ciblée.

Sur les 350 réponses de notre questionnaire, 58.3% disaient ne pas participer aux activités culturelles de tels organismes étudiants, contre 41.1% qui disaient y participer. Ainsi, si l'on s'en suit à l'hypothèse que l'offre proposée par les écoles et les universités est adaptée aux envies des étudiants et que celle-ci leur est rendue accessible, alors moins de la moitié de notre échantillon porte de l'intérêt à la Culture. Il est très

---

6. D'après une enquête réalisée par l'institut de sondages d'opinion 'Louis Harris 2' en 2014.

fastidieux en économie d'expliquer les choix des agents. La participation des étudiants aux activités culturelles ne dépendrait pas uniquement d'un arbitrage rationnel fonction du coût ou de la proximité de l'évènement par exemple, mais dépendrait alors aussi de facteurs psychocognitifs qui déterminent les envies.

## 2. Les goûts et habitudes

Les études sociologiques sur le sujet ont montré que la classe sociale dont une personne est issue est déterminante des habitudes de 'consommation culturelle'. Ainsi, la sensibilisation à un type de culture durant l'enfance joue sur l'attrait que portera l'individu à certaines pratiques dans sa future vie d'adulte. Un enfant habitué dès le plus jeune âge à aller au théâtre ou à l'opéra sera plus enclin à reproduire ces habitudes dans le futur. En outre, nous pensons que les pratiques culturelles des jeunes sont déterminées en grande partie par l'habitus, défini par Durkheim comme "l'ensemble des apprentissages, des dispositions acquises par l'enfant au cours de son éducation" ou comme "la capacité d'engendrer des pratiques" selon Marcel Mauss.<sup>7</sup> Ainsi, l'habitus englobe l'ensemble des conduites, des perceptions et des jugements transmis au cours de la socialisation, qui sont propres à une culture, à une histoire.

Deux variables illustrent alors ce principe dans notre enquête : la question de sensibilisation à la culture qui peut venir à la fois des parents et des cours extrascolaires, mais également la CSP d'appartenance du père. En effet, les études en sciences sociales et notamment celles de Bourdieu sur les héritiers, montrent que les habitudes de pratiques culturelles sont différentes selon le milieu social d'appartenance - même si l'école tend à réduire ces inégalités. Bien que la catégorie socio-professionnelle des parents soit une variable indépendante du sujet traité et donc un déterminant intrinsèque aux individus, nous l'introduisons dès cette partie puisque son lien à la variable de sensibilisation culturelle est fort, comme nous venons de le montrer. Nous savons également que les différences de pratiques entre les milieux sociaux ne sont pas uniquement dues aux habitudes mais également à l'accessibilité en termes d'argent et d'information. C'est ce que nous verrons dans la partie qui suit.

## 3. L'accessibilité

Une des questions qui nous semblait importante lorsque nous avons commencé cette enquête, était de savoir si chaque étudiant avait la possibilité de pratiquer ou d'avoir accès à la forme d'art qu'il affectionnait. Ainsi, nous avons demandé aux étudiants s'il y avait un écart entre leurs pratiques culturelles **effectives**, et celles qu'ils **souhaiteraient** avoir. Deux variables concernent ce point dans notre questionnaire.

Dans un premier temps nous avons sondé l'intérêt que portait chaque répondant aux activités dont nous souhaitions expliquer la fréquence : les expositions, les concerts, les bibliothèques etc. Trois modalités étaient possibles ; "ça ne m'intéresse pas", "pourquoi pas" ou encore "j'aimerais beaucoup". L'objectif était de déterminer l'appétence des étudiants à chaque activité culturelle, reflétant son ouverture à la Culture. Ensuite, afin de mesurer leur sentiment global et comprendre s'il y avait un décalage entre leurs désirs et leurs pratiques effectives, nous avons demandé s'ils souhaitaient améliorer/diversifier leurs pratiques culturelles, ou y consacrer plus de temps. À cette question, 90% de notre échantillon a répondu par l'affirmatif, dont la moitié souhaitait diversifier ses pratiques et l'autre moitié souhaitait consacrer plus de temps aux pratiques actuelles.

---

7. DIDIER, "Habitus".

L'accessibilité semble alors être une composante essentielle de la pratique d'une activité culturelle. Elle se manifeste sous plusieurs formes et son absence peut constituer un frein aux pratiques culturelles. Tout d'abord le coût des activités est considéré comme un des principaux freins aux sorties culturelles. En effet, même si les étudiants bénéficient de tarifs avantageux, certaines activités comme le cinéma par exemple peuvent avoir un coût non négligeable sur le budget. Aussi, pour 46.3% d'entre eux le manque d'information est également un frein, tout comme l'accessibilité géographique et la proximité des lieux d'expositions/de représentation pour 28.9% de nos répondants. Vient ensuite le sentiment que certaines activités ne sont pas démocratisées et semblent être réservées à une certaine catégorie de personne (23% de l'échantillon éprouve ce sentiment). De plus, 73% d'entre eux considèrent également manquer de temps.

Ainsi, pour 62% des étudiants le coût des activités est un des freins principaux à la pratique culturelle. C'est donc un point essentiel dans l'explication des fréquences. Nous avons cherché à le préciser en incluant dans le questionnaire une variable "budget" constituée de classes de dépenses dans les sorties culturelles pas mois (la variable concerne donc uniquement les **sorties** et non les activités du quotidien).

### 2.2.2 Les caractéristiques des individus

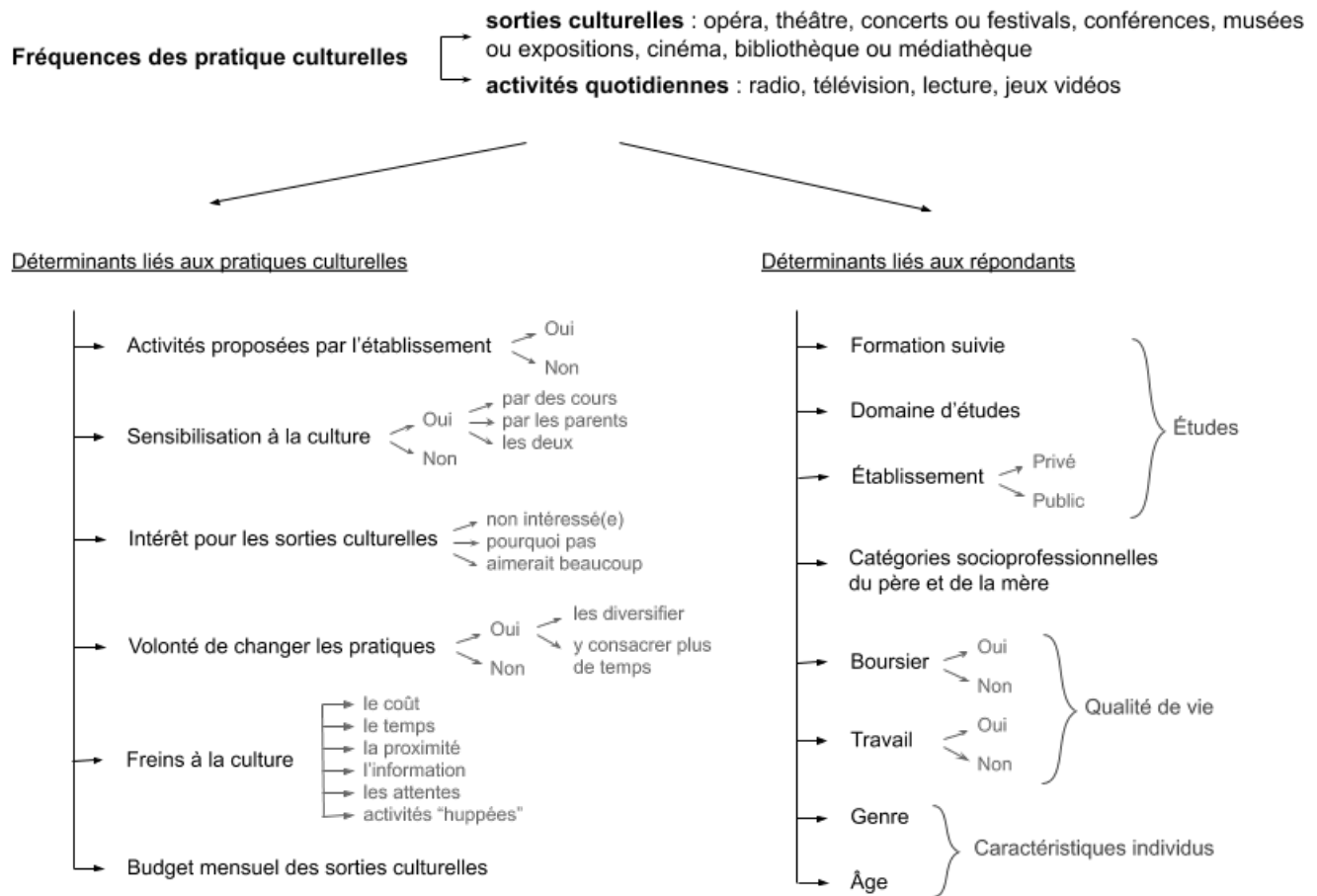
Dans un second temps, nous nous sommes penchées sur les caractéristiques propres à chaque répondant afin de pouvoir établir les profils des individus, indépendamment de l'intérêt personnel qu'ils portent à la culture.

Des questions sur la formation suivie, la nature de l'établissement (public ou privé), ainsi que le domaine d'études peuvent ainsi nous aider à comprendre l'intensité ou l'absence de pratiques culturelles chez certains étudiants. En effet, un étudiant en classe préparatoire ayant beaucoup de travail consacrera moins de temps aux activités culturelles qu'un étudiant en histoire ou en sociologie pour lequel de telles activités viennent simplement compléter son cursus (c'est le cas par exemple des musées, des expositions, des conférences etc.). De la même manière, comme nous l'avons évoqué précédemment, la catégorie socio-professionnelle des parents peut avoir une influence sur l'*habitus* transmis au répondant - apprentissages qui détermine en grande partie ses habitudes et ses manières d'être en société.

Pour revenir sur l'aspect onéreux des sorties culturelles qui peut freiner la majeure partie des étudiants, nous avons inclus 2 variables qui visent à identifier si le répondant est aisé ou s'il rencontre des difficultés financières pendant ses études. Il s'agit des variables 'boursier' et 'travail' qui prennent les modalités 1 si l'étudiant travaille ou s'il touche une bourse, 0 sinon. Finalement, nous avons demandé l'âge ainsi que le sexe de la personne pour voir s'il existe des disparités selon le genre du répondant et son âge ; nous pouvons par exemple supposer qu'un jeune étudiant venant d'obtenir son baccalauréat sera moins intéressé par des événements culturels qu'un étudiant en fin de licence ou en master qui a acquis une certaine maturité pendant son cursus. Ou au contraire ; puisqu'au fur et à mesure que l'on avance dans les études le travail s'intensifie, un jeune étudiant aura plus de temps à consacrer aux pratiques culturelles.

Nous aurons l'occasion de vérifier ces hypothèses dans une nouvelle partie, mais regardons avant cela un schéma récapitulatif des variables de notre questionnaire, à partir duquel nous avons construit notre base de données.

FIGURE 2 – Les variables issues de notre questionnaire



## 3 Analyse préliminaire

### 3.1 Rappels sur le nettoyage de la base

Rappelons rapidement le nettoyage que nous avons effectué l'année dernière pour la première analyse sur la base issue de notre questionnaire. Nous avons procédé au recodage de 3 variables :

- **Formation suivie** ; nous avons regroupé "écoles de commerce" et "écoles d'ingénieurs" en une seule modalité "École" qui compte 26 observations.
- **Domaine d'études** ; des réponses "autres" étaient nombreuses et concernaient des branches d'études de l'économie, nous avons donc créé les modalités "Gestion" avec 37 observations et "Finances" avec 9 observations, mais avons souligné le biais possible lié au manque à gagner des observations dans ces modalités lors de la diffusion du questionnaire.
- **Freins à la culture** ; 6 réponses "autres" concernant le manque d'envie ont été regroupées en une modalité portant ce nom.

Nous ajoutons à cela le recodage de la question portant sur la volonté de changer ou d'améliorer ses pratiques culturelles. Effectivement, comme visible en figure n°2, une question binaire demandait aux étudiants s'ils voulaient modifier leurs pratiques culturelles et une autre demandait s'ils souhaitaient y consacrer plus de temps ou les diversifier. Nous avons donc dans cette seconde variable des valeurs manquantes pour les personnes qui ne voulaient pas modifier leurs habitudes culturelles ; nous avons alors regroupé ces modalités en une seule variable prenant la valeur 0 si le répondant ne souhaite pas modifier ses pratiques, 1 s'il veut y consacrer plus de temps et 2 s'il veut les diversifier.

De plus, nous avons procédé à la randomisation de la variable "budget" qui était jusqu'ici composée de classes de dépenses mensuelles pour les activités culturelles. En effet, nous allons dans les parties suivantes, construire des modèles additifs généralisés visant à modéliser des relations non linéaires entre les variables à expliquer et certains facteurs explicatifs. Lors de notre première analyse l'année passée, l'utilisation d'un modèle multinomial ordonné sur la fréquence du théâtre avait révélé des effets de seuils de la variable budget. Jusqu'à la tranche 51-70 euros, la fréquentation culturelle était croissante du budget. En revanche, après 70 euros, plus le budget augmentait, plus la fréquence de pratiques culturelles baissait. Nous soupçonnons ainsi un lien non linéaire entre la fréquentation culturelle et le budget. Pour pouvoir l'étudier avec différentes clés, nous avons randomisé la variable du budget en faisant le choix arbitraire de supposer que sa distribution est aléatoire au sein des modalités. Pour ce faire, nous avons construit un algorithme sur VBA qui attribue pour chaque classe de budget, un nombre compris entre les bornes inférieures et supérieures afin que la distribution soit proche de celle de la variable initiale. Ainsi, si un individu appartient à la classe [1-10] qui correspond à la modalité n°2 de la variable initiale, un nombre entre 1 et 10 compris lui sera attribué aléatoirement - représentant son budget alloué aux pratiques culturelles. Nous pouvons voir en figure n°3 ci-dessous, le code sous VBA qui a servi à rendre le budget aléatoire.

Nous appelons la variable quantitative ainsi créée "budget\_random" et nous l'intégrerons dans les différentes modélisations. Les valeurs prises par cette dernière s'étendent de 0 à 117€ avec une médiane de 19€ et une moyenne de 24.84€. Nous voyons alors que certaines valeurs tirent la moyenne vers le haut ; comme sur toute variable quantitative, il convient désormais de vérifier ses points atypiques et sa corrélation avec notre autre variable quantitative qu'est l'âge des étudiants. Pour cette dernière, nous avons déjà retiré 8 observations l'année passée, considérées comme atypiques à partir du moment où l'âge excédait 28 ans.

FIGURE 3 – Programme sous VBA pour randomiser la variable budget

```

Sub budget_aleatoire()
    Cellule = ActiveCell
    Randomize

    Set Plage = Range(ActiveCell, ActiveCell.End(xlDown))

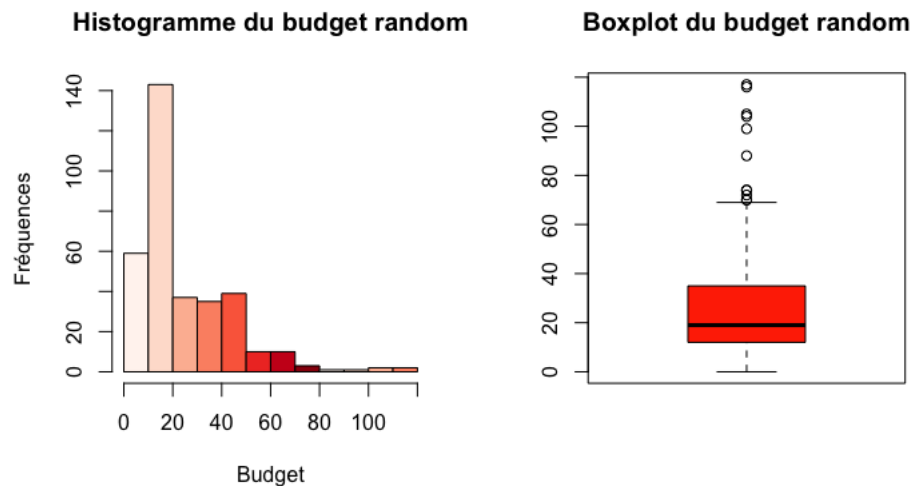
    For Each Cellule In Plage
        Cellule.Offset(0, 1) = "budget randomisé"

        Select Case Cellule.Value
            Case Is = 0: Cellule.Offset(0, 1) = 0
            Case Is = 1: Cellule.Offset(0, 1) = Int((10 - 1 + 1) * Rnd + 1)
            Case Is = 2: Cellule.Offset(0, 1) = Int((20 - 11 + 1) * Rnd + 11)
            Case Is = 3: Cellule.Offset(0, 1) = Int((50 - 21 + 1) * Rnd + 21)
            Case Is = 4: Cellule.Offset(0, 1) = Int((70 - 51 + 1) * Rnd + 51)
            Case Is = 5: Cellule.Offset(0, 1) = Int((100 - 71 + 1) * Rnd + 71)
            Case Is = 6: Cellule.Offset(0, 1) = Int((130 - 100 + 1) * Rnd + 100)
        End Select

        ActiveCell.Offset(0, 1).Select
    Next
End Sub

```

FIGURE 4 – Histogramme et boxplot du budget randomisé



D'après la figure n°4 nous voyons que la majorité des valeurs prises par la variable du budget se situent entre 10 et 20€ de dépenses par mois pour les sorties culturelles. Nous constatons aussi que certaines valeurs sont très élevées, elles se situent en queue de distribution avec un budget mensuel supérieur à 100€. Nous retrouvons ces valeurs extrêmes sur le boxplot puisqu'elles sont hors de la boîte à moustaches : elles dépassent le 3ème quantile. Visuellement, 9 valeurs semblent ainsi atypiques, voyons ce que le test statistique nous dira.

TABLE 1 – Résultats du test de Rosner sur la variable "budget\_random"

Outliers détectés visuellement	Seuil d'atypicité du test	Outliers détectés statistiquement	Observations concernées
9	$\geq 88\text{€}$	6	n° 340, 2, 197, 302, 122 et 116

Nous voyons d'après le tableau n°1 que seulement 6 observations sur 9 potentiellement atypiques sont finalement à retirer de la base. Effectivement, le budget ainsi randomisé est considéré comme atypique à partir du moment où l'étudiant dépense plus de 87€ pour les activités culturelles par mois. Nous retirons donc ces 6 individus et obtenons une base aux dimensions 336 x 33. L'histogramme de distribution de la variable sans les points atypiques se trouve annexe n°1 ; la queue de distribution que nous observions jusqu'alors est maintenant inexistante.

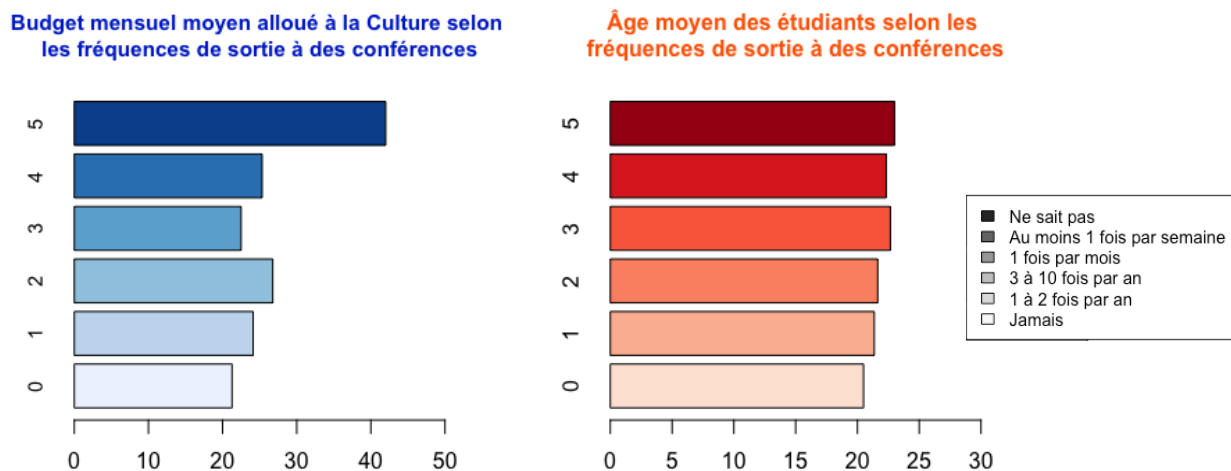
Concernant la corrélation qui lie nos 2 variables quantitatives nous voyons sur la matrice de corrélation en annexe n°2 que le coefficient est de 0.072 ; il n'existe donc pas de corrélation trop élevée qui puisse altérer les résultats de nos futures analyses. Nous pourrions inclure à la fois l'âge et le budget randomisé dans nos modèles. De même, de manière à ne pas inclure de variables dépendantes dans les modèles, nous avons l'année dernière appliqué le test de Chi-2 aux variables qualitatives. 5 types de variables étaient alors à exclure des différentes modélisations ; la fréquence de sortie à l'opéra, les freins aux pratiques culturelles, la CSP d'appartenance de la mère, et la nature de l'établissement (public ou privé). De plus, chaque variable d'intérêt pour une activité était corrélée avec la variable de fréquence de cette même activité. Ainsi, la variable d'intérêt pour le théâtre est corrélée à la variable de fréquentation du théâtre, ce qui nous semble logique puisque plus l'on porte d'intérêt à une activité plus l'on sera enclin à la pratiquer.

Ainsi nettoyée, nous disposons d'une base de données composée de 336 observations et de 33 variables : soit le même nombre que l'année dernière (puisque l'on a regroupé les changements de pratiques culturelles en une, et avons créé une variable budget randomisé), et 6 observations de moins (les points atypiques de cette dernière). Il est important de vérifier la répartition des répondants dans les modalités pour voir s'il n'y a pas de biais lié à la sur ou sous représentation d'une population au sein de notre échantillon. Les fréquences de pratique des activités culturelles proposées en première question de notre enquête diffèrent grandement selon le type d'activité. Ainsi, près de 90% de notre échantillon ne se rend jamais à l'opéra (taux calculé sur la base nettoyée), tandis que 67% regarde la télévision au moins une fois par semaine. Ces chiffres ne sont pas étonnants mais ils ne nous permettent pas de réaliser une analyse en gardant les 5 modalités distinctes ; il faudrait alors procéder à un recodage des variables en regroupant les modalités entre elles de manière à se rapprocher le plus de l'équi-répartition dont nous parlions en partie précédente. De la même manière, nous constatons que 91% des répondants étudient dans un établissement public ; le biais est lié au fait que la grande majorité des réponses aient été obtenues via la mail-liste de diffusion de la faculté d'économie. Il apparaît aussi que 67.6% des individus sont des filles ; peut-être sont-elles plus intéressées par la culture en général, ce qui les a amenées à répondre davantage à notre questionnaire. Nous savons également que les femmes répondent globalement davantage aux questionnaires que les hommes. Quoi qu'il en soit, il sera intéressant de croiser la variable "genre" avec certaines fréquences de pratiques culturelles dans le but de voir si cette dernière a vraiment un impact sur les sorties culturelles. Enfin, comme nous l'avons souligné l'année dernière dans notre première analyse, 70% de notre échantillon est en étude d'économie (ou ses dérivés) puisque la mail-liste était notre principal canal de diffusion.

### 3.2 Quelques statistiques descriptives

Comme énoncé précédemment, les variables à expliquer doivent être équiréparties entre la pratique de l'activité et la non-pratique, de manière à avoir une analyse viable. Pour cette raison, nous expliquerons exclusivement la probabilité de fréquentation du théâtre et celle des conférences de manière simultanée dans un modèle probit bivarié. Avant de commencer l'analyse économétrique à proprement parler, regardons quelques statistiques qui peuvent nous éclairer ou nous donner des informations pour les futures modélisations.

FIGURE 5 – Âge et budget moyens selon les fréquences de sortie à des conférences

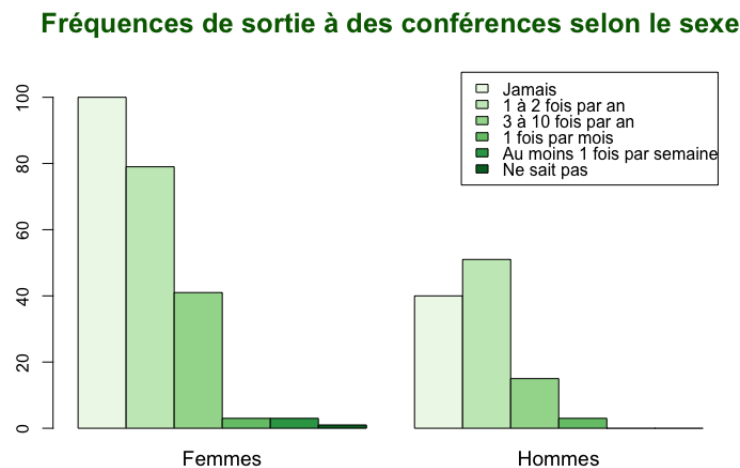


Les barplots ci-dessus et les tableaux croisés disponibles en annexe n°3 nous informent qu'en moyenne l'âge n'a pas un impact très important sur la fréquentation de conférences puisque les barres correspondant aux fréquences de sorties sont plus ou moins égales selon l'âge du répondant. On constate cependant un léger crescendo dans la fréquentation des conférences à mesure que l'âge de l'individu augmente. Ainsi, ceux qui ne se rendent jamais à des conférences ont un âge moyen entre 20 et 21 ans (20.49 en annexe), tandis que les personnes s'y rendant 1 fois par mois ont en moyenne entre 22 et 23 ans.

Concernant le budget (nous avons évidemment pris la variable randomisée) le constat est différent ; on voit d'emblée que les répondants ne sachant pas à quelle fréquence ils se rendent à des conférences (modalité n°5) sont ceux qui allouent le budget le plus élevé aux sorties culturelles (42€ en moyenne contre 22.5€ pour ceux qui s'y rendent une fois par mois). On peut alors se demander s'ils ont répondu correctement au questionnaire ou s'ils n'avaient idée ni du budget ni des fréquences de leurs pratiques culturelles.



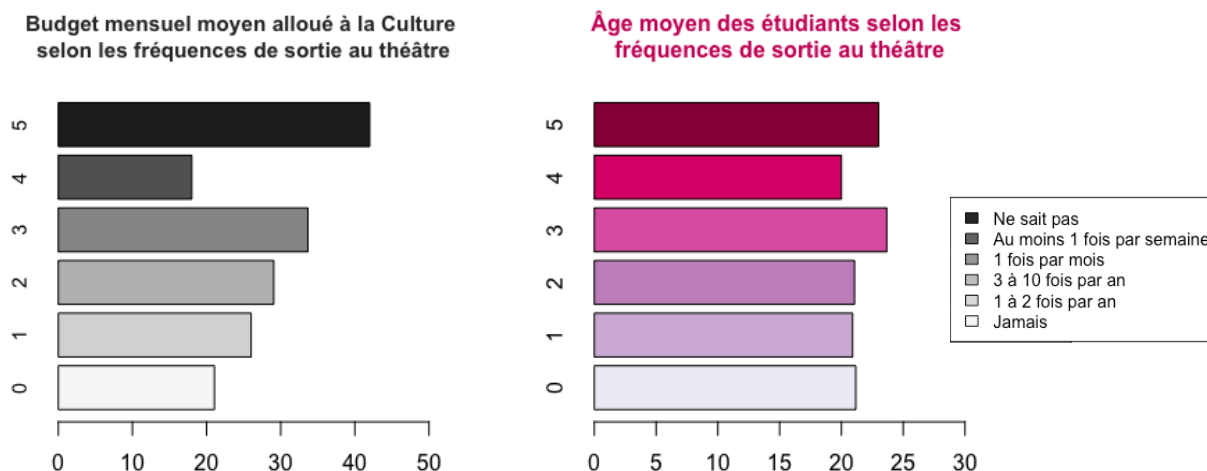
FIGURE 6 – Les fréquences de sortie à des conférences selon le genre du répondant



Le graphique 6 ci-dessus nous permet de constater visuellement que notre base contient plus de réponses féminines que masculines - comme nous avons pu le souligner précédemment. En nous aidant des effectifs des tableaux croisés disponibles eux aussi en annexe n°3, nous voyons que 44% d'entre elles ne se rendent jamais à des conférences tandis que ce taux est de seulement 37% chez les hommes. En outre, étant donné que la part des personnes ayant répondu "ne sait pas" est négligeable, nous remarquons qu'en proportion les hommes se rendent davantage à des conférences que les femmes sur notre échantillon.

Sachant que nos modélisations porteront sur les fréquences de sorties à des conférences **et** au théâtre nous réitérons notre analyse sur cette dernière variable.

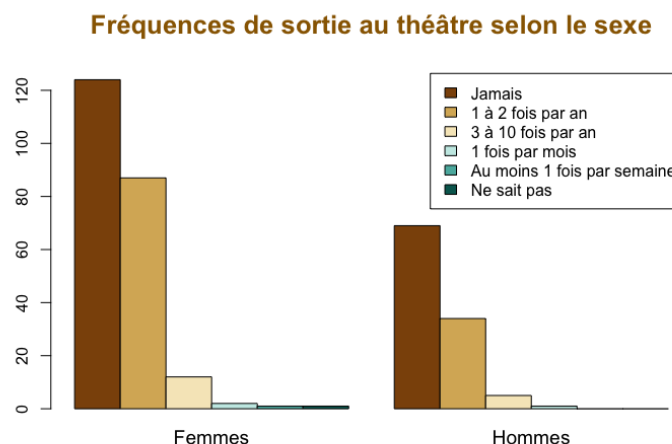
FIGURE 7 – Âge et budget moyens selon les fréquences de sortie au théâtre



On constate cette fois une plus grande disparité dans les âges moyens selon les fréquences de sortie au théâtre plutôt qu'à des conférences. En effet, on voit sur la figure n°7 ainsi que l'annexe n°4 que l'âge moyen

des répondants qui vont 1 fois par moi au théâtre (modalité 3) se situe entre 23 et 24 ans tandis qu'il était entre 22 et 23 ans pour les conférences, et qu'il est de 20 ans pour les étudiants se rendant au théâtre au moins une fois par semaine. Aussi, la moyenne d'âge des individus se rendant au théâtre est de 2 à 3 ans plus faible que celle des individus se rendant à des conférences au moins une fois par semaine. De même, si l'on s'intéresse au budget moyen alloué aux sorties culturelles on voit que pour les mêmes fréquences de sorties entre ces 2 activités (théâtre et conférences) les étudiants se rendant au théâtre dépensent davantage que ceux qui vont à des conférences (à fréquences égales). Cela peut s'expliquer par un coût à l'entrée plus élevé pour les pièces de théâtre que pour des conférences plus souvent accessibles gratuitement.

FIGURE 8 – Les fréquences de sortie au théâtre selon le genre du répondant



Comparé aux fréquences de sorties à des conférences, on voit en figure n°8 et en annexe n°4 que les étudiants n'allant jamais au théâtre sont plus nombreux : cela représente 54.6% des femmes et 63.3% des hommes contre respectivement 44 et 36.7% pour les conférences. Il apparaît ainsi que les sorties au théâtre sont plus rares que les sorties à des conférences chez les étudiants nantais de notre échantillon.

Pour résumer, nous avons constaté que l'âge est fonction croissante des sorties aux conférences comme nous avons pu le voir en figure n°5. De même, les étudiants allouant le budget le plus élevé aux sorties culturelles par mois sont aussi ceux qui ne connaissent pas leurs fréquences de sorties - que ce soit à des conférences ou au théâtre. De manière générale, les individus qui ne se rendent jamais à de telles activités sont ceux qui dépensent le moins dans les sorties culturelles, ce qui paraît logique dans ce raisonnement. Enfin, en proportion il apparaît sur notre échantillon que les hommes se rendent d'avantage au théâtre que les femmes (moins d'individus dans la modalité "jamais") mais le constat est différent en ce qui concerne les conférences puisqu'alors la proportion de femmes s'y rendant dépasse celle des hommes. Nous pouvons retrouver effectivement ces taux dans le tableau qui suit.

TABLE 2 – Part des individus ne se rendant jamais aux activités culturelles suivantes

	Femmes	Hommes
Théâtre	54.6%	63.3%
Conférences	44.1%	36.7%

### 3.3 Arbres de décision - CART

Les arbres de décisions ('Classification And Regression Trees' d'après l'acronyme anglais), constituent une méthode supervisée de Machine Learning (ML) qui peut être utilisée dans la classification. En ML, un arbre est une classe d'algorithmes non paramétriques qui fonctionnent en partitionnant l'espace d'entités en un certain nombre de régions plus petites avec des valeurs de réponse similaires, à l'aide d'un ensemble de règles de fractionnement. Son fonctionnement est le suivant : à chaque noeud, l'algorithme considère les  $N$  variables et cherche celle qui divise le jeu de données de la manière la plus optimale possible en maximisant la diminution globale de l'impureté, c'est-à-dire en réduisant le plus possible l'erreur. Plus simplement, à chaque noeud l'algorithme considère les variables les plus influentes et divise à partir de celles-ci un ensemble en 2 groupes qui sont les plus hétérogènes entre eux, mais les plus homogènes en leur sein.

L'application de CART sur notre jeu de données peut nous permettre d'avoir un premier aperçu sur les variables qui divisent le mieux le jeu, et aussi celles qui sont les plus pertinentes dans l'explication des variables à expliquer. Ainsi nous commençons par construire un arbre sur la fréquentation des théâtres puis sur celle des conférences, phénomènes que nous chercherons à modéliser dans les parties suivantes.

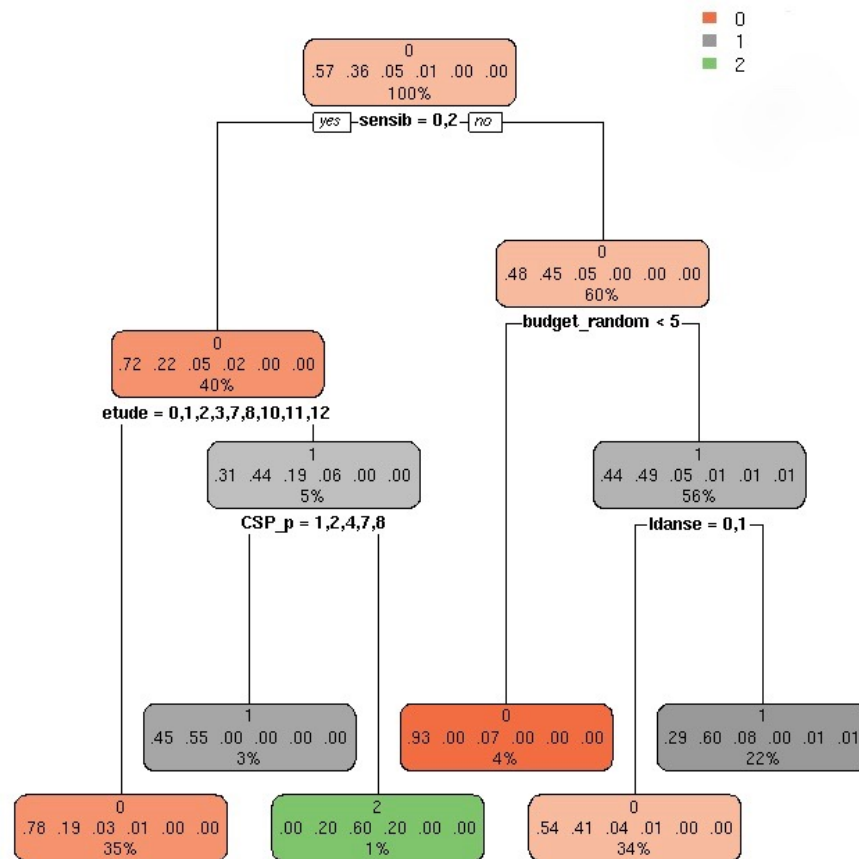


FIGURE 9 – Arbre de décision sur la fréquentation des théâtres

L'arbre nous permet d'avoir une première idée sur l'importance de chaque variable et des liens qui les relient entre elles. On voit ainsi que l'arbre part de la variable de sensibilisation à la Culture pour dichotomiser le jeu de données en 2 groupes homogènes : il sélectionne les modalités 0 et 2 de cette variable, à savoir ne pas

avoir été sensibilisé du tout et avoir été sensibilisé par des cours extra-scolaires. Puis, pour ceux qui ne font pas partie des modalités 0 et 2 de cette variable (qui ont donc été sensibilisés à la culture par leurs parents ou par leurs parents ET des cours extra-scolaires), l'arbre segmente l'échantillon et on voit que 60% allouent un budget mensuel moyen de moins de 5€ aux sorties culturelles. En revanche, parmi les répondants appartenant aux modalités 0 et 2 de la variable "sensib", 40% étudient l'économie, la gestion, la finance, la sociologie, la psychologie, la philosophie, l'architecture ou les sciences. L'arbre montre aussi que les étudiants déboursant moins de 5€ par mois pour des activités culturelles ont un intérêt pour la danse limité (modalités 0 et 1 de la variable "Idanse"). On voit donc que les variables qui divisent au mieux les étudiants sont celles de l'arbre : la sensibilisation à la culture, le domaine d'étude, le budget randomisé, la catégorie socio-professionnelle du père et l'intérêt pour la danse.

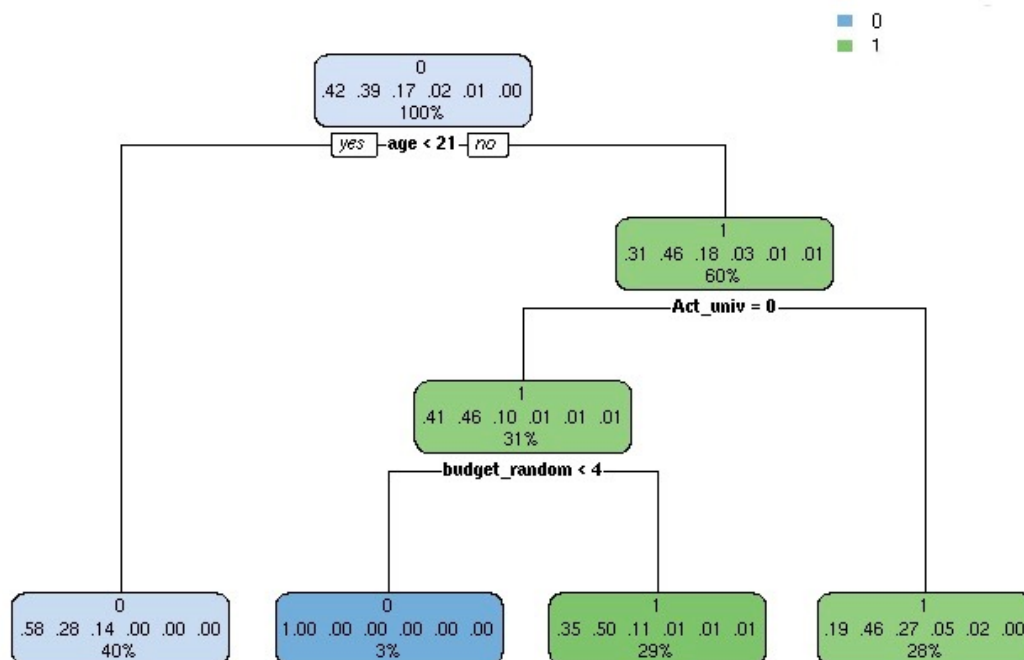


FIGURE 10 – Arbre de décision sur la fréquentation des conférences

Sur la figure n°10 on voit que pour la fréquentation des conférences, il y a à peu près autant d'étudiants qui ont moins de 21 ans (40%) que de plus de 21 ans (60%) dans notre base. De plus, 28% des individus de plus de 21 ans participent aux activités culturelles proposées par leur établissement (modalité 1 de "Act\_univ"). Parmi les personnes qui semblent peu attirées par les activités culturelles (cf. modalité 0 de la variable "act\_univ" qui englobe 31% des étudiants de plus de 21 ans), près de 10%<sup>8</sup> dépensent moins de 4€ par mois pour de telles activités. Attention ici à l'interprétation puisqu'initialement, la variable budget était constituée de classes donc l'arbre aurait été peut être différent en prenant la variable d'origine. Dans tous les cas, l'arbre nous dévoile que dans l'explication des fréquences de sorties aux conférences, la variable qui segmente au mieux l'échantillon est celle de l'âge. Pour le théâtre comme pour les conférences, nous nous attendons à ce que les variables dichotomisant le jeu au premier et deuxième noeuds, soient significatives dans l'explication de la pratique de ces activités.

8.  $0.03/0.31 \times 100$

## 4 Modèle biprobit

### 4.1 Fonctionnement théorique d'un modèle biprobit

Le modèle probit bivarié sert à expliquer la probabilité de deux événements simultanés, il est ainsi composé de 2 équations et s'applique lorsque l'on cherche à modéliser simultanément deux variables qualitatives dichotomiques. Il se présente de la manière suivante :

$$Y_{i1} = x_{i1}\beta_1 + \epsilon_{i1} \quad \text{où} \quad Y_{i1} = 1 \quad \text{si} \quad Y_{i1} > 0$$

$$Y_{i2} = x_{i2}\beta_2 + \epsilon_{i2} \quad \text{où} \quad Y_{i2} = 1 \quad \text{si} \quad Y_{i2} > 0$$

Les variables  $x_i$  sont des combinaisons linéaires pouvant être les mêmes pour chacune des équations du biprobit, ou différentes selon le phénomène expliqué.

L'estimation du modèle probit bivarié se fait sous les hypothèses des termes d'erreurs  $\epsilon_i$  suivantes :

- La loi conjointe des termes  $\epsilon_{i1}$  et  $\epsilon_{i2}$  est normalement distribuée
- Leur espérance mathématique est nulle, telle que  $E(\epsilon_{i1})=E(\epsilon_{i2})=0$
- Leur variance est normalisée à 1, telle que  $\text{Var}(\epsilon_{i1})=\text{Var}(\epsilon_{i2})=1$
- Leur covariance est égale à  $\rho$  qui est le coefficient de corrélation entre les termes d'erreurs, telle que  $\text{Cov}(\epsilon_{i1}, \epsilon_{i2})=\rho$

Les estimateurs du biprobit s'obtiennent par la maximisation numérique du log de vraisemblance. Dans le cas où les termes  $\epsilon_{i1}$  et  $\epsilon_{i2}$  ne sont pas corrélés ( $\rho$  non significatif), la densité bivariée  $\phi_2$  est égale au produit des densités marginales.

Appliquons à présent ce modèle sur notre base, et plus précisément sur le fait de se rendre conjointement au théâtre et à des conférences.

### 4.2 Application sur les fréquentations du théâtre et des conférences

Ayant analysé les fréquences de sortie au théâtre l'année précédente, nous avons décidé de compléter cette étude en modélisant conjointement les sorties au théâtre avec les sorties à des conférences, puisque nous pensons qu'il existe un lien entre ces 2 activités culturelles. De même, il s'agit là des variables de fréquence les mieux réparties de notre enquête. Effectivement, si nous les recodons de manière à obtenir  $Y=1$  si l'étudiant se rend à l'activité au moins une à deux fois par an et  $Y=0$  sinon, nous obtenons un partage proche de l'équi-répartition. Nous retrouvons ainsi en  $Y=1$  les modalités 1, 2, 3, 4 et 5 qui correspondent respectivement à des sorties 1 à 2 fois par an, 3 à 10 fois par an, 1 fois par mois, au moins 1 fois par semaine et 'ne sait pas'. Nous considérons la modalité "ne sait pas" dans la pratique de l'activité puisque nous supposons que si l'étudiant ne pratiquait jamais telle ou telle activité, il n'hésiterait pas sur sa fréquence.  $Y=0$  englobe quant à lui la modalité 0 des fréquences qui correspond à la non-pratique de l'activité en question (modalité "jamais").

Nous retenons ainsi  $Y_1$  le fait de se rendre au théâtre ou non, et  $Y_2$  le fait de se rendre à des conférences ou non - et nous regardons la répartition des répondants au sein de ces 4 probabilités.

TABLE 3 – Répartition des individus dans les modalités

	$Y_2=0$	$Y_2=1$	Total
$Y_1=0$	91 (27%)	102 (30%)	193 (57%)
$Y_1=1$	49 (15%)	94 (28%)	143 (43%)
Total	140 (42%)	196 (58%)	336 (100%)

D'après la table n°3 on voit que les étudiants nantais ayant répondu à notre questionnaire sont bien répartis dans les modalités ; la répartition parfaite voudrait qu'il y ait 25% des étudiants dans chaque probabilité. Nous voyons par ailleurs que la modalité contenant le moins de répondants correspond à une personne qui se rend au théâtre même occasionnellement ( $Y_1=1$ ) mais jamais à des conférences ( $Y_2=0$ ), où se trouve seulement 15% de notre échantillon. Cependant, celle qui regroupe le plus d'étudiants - pour un total de 102 répondants c'est à dire 30% de notre échantillon, correspond à la situation inverse de celle qui regroupe le moins d'étudiants : se rendre à des conférences mais pas au théâtre. De même, nous remarquons que les moyennes de chaque  $Y$  se situent autour de 50% (à 10% près) ce qui prouve une fois encore la bonne répartition des individus dans les différentes classes. Ainsi, 57% des étudiants de notre échantillon se rendent à des conférences et 42% se rendent au théâtre.

Nous pouvons maintenant modéliser un biprobit qui lie ces 2 activités culturelles avec l'hypothèse qu'il existe une certaine dépendance entre les phénomènes. De plus, nous supposons que la matrice de variance-covariance dans le terme d'erreur  $\rho$  suit une loi normale, voilà pourquoi nous modélisons par un biprobit et non un bilogit.  $\rho$  modélise le lien entre les 2 modèles probits ; ainsi, s'il est significatif modéliser les phénomènes conjointement était utile, dans le cas contraire cela signifie que les erreurs des 2 modèles ne sont pas liées et donc qu'il convient de faire 2 modélisations distinctes. Lors des rappels de nettoyage de la base nous avons vu que certaines variables étaient dépendantes entre elles et que par conséquent il fallait rester vigilants sur les estimations pour éviter tout problème de liaisons qui viendrait les fausser. Par rapport au dossier de l'année passée, l'erreur à ne pas reproduire consiste à ajouter les fréquences des autres pratiques culturelles comme explicatives de la fréquentation du théâtre et des conférences. Étant donné que nous disposons de 21 variables explicatives (hors fréquences et en considérant seulement le budget randomisé), nous allons procéder à une sélection de variables pour ne retenir que les plus pertinentes. Nous appliquons alors les méthodes backward, forward et stepwise pour  $Y_1$  et  $Y_2$  et incluons les variables sélectionnées dans le modèle biprobit que nous exécuterons sous le logiciel Stata.

TABLE 4 – Sélection de variables pour les  $Y_i$ 

$Y_1$ : Théâtre	$Y_2$ : Conférences
Intérêt pour l'opéra	Intérêt pour les bibliothèques
Intérêt pour les musées	Intérêt pour la musique classique
Budget alloué aux sorties culturelles	Participation aux activités de l'établissement
Volonté de changer ses pratiques	Âge
Sensibilisation à la Culture	

On peut voir dans la table n°4 les résultats des méthodes de sélection de variables forward, backward

et stepwise qui donnent les mêmes résultats pour les 2 pratiques culturelles à expliquer. Nous savons ainsi quelles variables retenir pour chaque Y du modèle biprobit ; il n'y a aucune variable en commun entre les 2 phénomènes à expliquer. Nous voyons de même, que la variable du budget rendu aléatoire est sélectionnée par les méthodes pour expliquer le fait de se rendre au théâtre mais pas pour les conférences. Cela apparaît comme logique puisque comme montré précédemment, les conférences sont plus souvent accessibles gratuitement tandis que les représentations au théâtre sont dans la plupart des cas payantes, ce qui constitue donc un frein à l'entrée.

Par rapport aux arbres de décisions, on voit que pour le théâtre les variables 'sensibilisation à la culture' et 'budget' que nous voyions importantes précédemment ont effectivement été sélectionnées. De même, pour les conférences, les variables 'âge' et 'participation aux activités proposées par l'établissement' ont été retenues. La différence de sélection dans les variables réside peut-être dans le fait que pour les arbres nous avons pris les variables initiales - c'est à dire composées de 5 modalités différentes, tandis que pour la sélection de variables nous avons pris les variables recodées en mode binaire (au vu des modélisations biprobit à venir).

TABLE 5 – Estimation du modèle biprobit sur la base nettoyée

	Variables	Coef	Err.Std	p-value
<b>Théâtre</b>	Intérêt pour l'opéra	0.352	0.097	0.000 ***
	Sensibilisation à la Culture	0.190	0.063	0.002 **
	Volonté de changer ses pratiques	-0.257	0.118	0.029 **
	Intérêt pour les musées	0.174	0.125	0.166
	Budget random alloué aux sorties culturelles	0.123	0.004	0.003 **
	Constante	-1.013	0.260	0.000 ***
<b>Conférences</b>	Participation aux activités de l'établissement	0.348	0.147	0.017 *
	Intérêt pour la musique classique	0.227	0.093	0.015 *
	Intérêt pour les bibliothèques	0.214	0.099	0.031 *
	Âge	0.162	0.039	0.000 ***
	Constante	-3.807	0.822	0.000 ***
	Nombres d'observations	336		
	Test de Wald	83.38		0.000 ***
	Test de $\rho=0$	2.046		0.153

D'après la table n°5 on constate d'emblée que le modèle biprobit n'est pas nécessaire. Effectivement, le paramètre qui lie les termes d'erreur des 2 modèles n'est pas significatif. La valeur associée au test du ratio de vraisemblance basé sur l'hypothèse nulle  $\rho=0$ , est de 0.153 ; nous acceptons ainsi  $H_0$ , il n'est pas pertinent de modéliser les phénomènes d'aller au théâtre et à des conférences conjointement. Par ailleurs, nous remarquons que 8 variables explicatives sur 9 sont significatives au seuil de risque de 5%, ce qui est assez satisfaisant par rapport aux sélections de variables appliquées précédemment. Il convient donc de construire 2 modèles séparés ; un qui explique le fait de se rendre au théâtre, et l'autre le fait de se rendre à des conférences.

### 4.3 Les modèles logit

Dans cette nouvelle section nous allons estimer deux modèles logit distincts des phénomènes que nous cherchons à expliquer. Nous passerons rapidement sur l'explication des hypothèses que doit valider un modèle logit puisque nous les avons étudiées en détail l'année dernière. Ainsi, nous modéliserons 2 modèles logit pour les fréquentations du théâtre et des conférences, et nous chercherons ceux qui conviennent le mieux, c'est à dire ceux qui valident les hypothèses et ont une bonne qualité sans détailler longuement. La démarche suivie pour trouver ces modèles se trouvera dans la partie des annexes.

#### 4.3.1 Le fait de se rendre à des conférences

Pour simplifier la démarche, nous avons, pour expliquer le fait de se rendre à des conférences ou non, supposé directement l'hétéroscédasticité des erreurs. Nous avons alors utilisé la fonction **hetglm()** sous R en mettant toutes les variables explicatives en potentiels coupables de l'hétéroscédasticité. Pour rappel, dans l'explication de la fréquentation de conférences nous incluons les variables sur la participation des étudiants aux activités proposées par leur établissement, leur intérêt pour la musique classique ainsi que pour les musées et leur âge en années. Comme visible en annexe n°5 dans un premier temps seule la variable sur la participation aux activités des établissements sort significative à 5% dans l'explication de l'hétéroscédasticité des erreurs. En réestimant un modèle où nous n'incluons que cette dernière comme coupable, nous constatons qu'elle n'est plus significative au seuil de risque de 5%. Les erreurs sont donc homoscedastiques et nous estimons un modèle logit contenant les 4  $X_i$  énoncés précédemment, grâce à la fonction **glm()**. Ce dernier valide toutes les hypothèses puisque :

- ✓ le **VIF** de chacune des variables se situe autour de 1 et est donc largement inférieur à 10
- ✓ le modèle a un **intérêt** ; la p-value associée au test de nullité des coefficients est de 1.44 e-08 et donc nous rejetons  $H_0$
- ✓ le modèle a une assez bonne **qualité d'ajustement** ; le  $R^2$  de Mac Fadden est de 0.11
- ✓ le modèle a une assez bonne **qualité prévisionnelle** ; le taux de prédiction juste est de 65.77% et donc supérieur à la moitié des observations
- ✓ il n'y a pas de **données influentes** ; tous les résidus des observations sont compris entre -2 et 2, sauf une qui se situe à -2.04 mais étant très proche de la borne -2, nous décidons de la laisser

Le modèle ainsi obtenu est visible en table n°6.

TABLE 6 – Estimation du modèle logit sur les conférences

Variables	Coef	Err.Std	p-value
Constante	-6.187	1.420	0.000 ***
Participation aux activités des établissements 1	0.740	0.251	0.003 **
Participation aux activités des établissements 2	-14.012	620.63	0.982
Intérêt pour la musique classique 1	0.871	0.292	0.003 **
Intérêt pour la musique classique 2	0.774	0.308	0.012 *
Intérêt pour les bibliothèques 1	0.630	0.323	0.051 .
Intérêt pour les bibliothèques 2	-0.786	0.337	0.020 *
Âge du répondant	0.243	0.066	0.000 ***
Critère AIC	422.42		



### 4.3.2 Le fait de se rendre au théâtre

Nous faisons de même pour la fréquentation du théâtre. Pour rappel nous disposons de 5 variables explicatives pour l'expliquer : l'intérêt pour l'opéra ainsi que pour les musées, le budget aléatoire alloué aux sorties culturelles, la volonté de changer ses pratiques et enfin le fait d'avoir été ou non sensibilisé à la culture pendant l'enfance. De la même manière que pour les conférences, nous commençons par estimer un modèle prenant en compte une éventuelle hétéroscédasticité des erreurs en mettant en potentiels coupables toutes les variables explicatives dont nous disposons pour ce modèle. Comme visible en annexe n°6, avec une première estimation, seules les variables de sensibilisation à la culture et du budget peuvent être coupables de l'hétéroscédasticité des erreurs. Nous modélisons une nouvelle fois en excluant les autres variables explicatives des potentiels coupables, et finalement, comme pour les conférences, aucune variable n'est significative. Nous basculons donc sur un modèle logit avec les erreurs homoscedastiques. Ce dernier contenant 4 données influentes que sont les 151, 59, 184 et 253, n'est pas valide. Nous retirons donc ces observations et réitérons notre analyse. Le modèle sortant auquel nous enlevons la variable non significative de l'intérêt pour les musées, est alors valide et pertinent :

- ✓ le **VIF** de chacune des variables se situe aussi autour de 1
- ✓ le modèle a un **intérêt** ; la p-value associée au test de nullité des coefficients est de 1.66 e-13
- ✓ le modèle a une assez bonne **qualité d'ajustement** ; le  $R^2$  de Mac Fadden est de 0.17
- ✓ le modèle a une assez bonne **qualité prévisionnelle** ; le taux de prédiction juste est de 70.18%
- ✓ il n'y a plus de **données influentes** ; les résidus supérieurs à 2.06 ont été retirés de la base

De la même manière que pour le logit sur le fait de se rendre à des conférences, nous pouvons regarder l'estimation du modèle sur le théâtre, disponible en table n°7.

TABLE 7 – Estimation du modèle logit sur le théâtre

Variables	Coef	Err.Std	p-value
Constante	-3.589	0.692	0.000 ***
Intérêt pour l'opéra 1	1.003	0.290	0.000 ***
Intérêt pour l'opéra 2	1.204	0.343	0.000 ***
Budget randomisé	0.026	0.008	0.001 ***
Volonté de changer ses pratiques 1	1.088	0.585	0.063 .
Volonté de changer ses pratiques 2	0.183	0.597	0.759
Sensibilisation à la culture 1	1.722	0.454	0.000 ***
Sensibilisation à la culture 2	0.998	0.469	0.033 *
Sensibilisation à la culture 3	1.815	0.423	0.000 ***
Critère AIC	392.07		

On a donc obtenu pour les fréquentations du théâtre et des conférences, 2 modèles logit distincts et qui valident les hypothèses. Notre objectif à présent, est d'estimer des modèles additifs généralisés contenant les mêmes variables et de voir si considérer la non-linéarité des liens permet d'améliorer significativement la qualité des modèles.

## 5 Modèle additif généralisé

Les régressions logistiques réalisées avec un logit sur les fréquences des activités culturelles a permis d'établir des liens linéaires entre les variables. Comme évoqué précédemment, nous cherchons dans cette partie à savoir si des liens non linéaires pourraient être plus appropriés pour représenter les relations entre les variables. Nous utiliserons pour cela les 2 variables quantitatives dont nous disposons, à savoir l'âge du répondant et le budget alloué aux sorties culturelles - que nous avons randomisé par un programme sous VBA. Pour étudier ces liens, nous appliquerons un modèle additif généralisé qui s'affranchit de la linéarité pour mieux prendre en compte les seuils et les valeurs extrêmes.

### 5.1 Fonctionnement théorique d'un modèle additif généralisé

Introduit dès les années 80, le modèle additif généralisé (GAM) permet de traiter des potentielles relations non linéaires que peuvent prendre les variables explicatives, en modélisant une variable à expliquer avec des fonctions de lissage non-paramétriques des prédicteurs. Ainsi, la variable étudiée n'est plus expliquée par des variables explicatives mais par des fonctions de ces dernières. Ainsi, l'espérance de  $Y_i$  ne dépend plus des valeurs que peuvent prendre les variables explicatives<sup>9</sup>. On obtient alors une relation telle que :

$$Y_i = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i})$$

où  $x_i$  sont des variables quantitatives. Ainsi la distribution de la variable ne suit plus une loi normale, mais celle-ci peut prendre la forme d'une loi binomiale, d'une loi de Poisson, de Gamma etc. Pour étudier le comportement de ces fonctions sur  $Y$ , il convient donc de définir et estimer les fonctions non linéaires  $f_j$  qui remplacent les coefficients  $\beta_j$ . Ainsi, l'effet additif de chaque prédicteur est conservé, tel que :  $Y = X\beta + \epsilon$  devient  $Y = f(X) + \epsilon$ . On obtient donc un modèle linéaire généralisé tel que :

$$g(E(Y_i)) = \beta_0 + \sum_{j=1}^p f_j(X_i^d)$$

On comprend ainsi que le modèle linéaire général est un cas particulier du modèle linéaire généralisé, où  $Y$  suit une loi normale et où les fonctions de liaisons avec les variables explicatives sont de simples fonctions identité telles que  $f(x)=x$ .<sup>10</sup> Les fonctions non spécifiques des variables prédictives permettent alors de maximiser la qualité de la prévision de  $Y$ .

### 5.2 Application sur la fréquentation du théâtre

De façon à prédire pour chaque étudiant la probabilité de la variable dichotomique fréquentation du théâtre/non-fréquentation, nous avons tout d'abord appliqué la méthode stepwise sur la modélisation logistique afin de connaître les variables les plus pertinentes à inclure dans notre modèle. Les variables dépendantes déterminées précédemment ont été exclues d'emblée. Nous obtenons logiquement les mêmes variables que pour pour le stepwise du modèle probit : l'intérêt pour l'opéra, la sensibilisation à la culture, la volonté de changer ses pratiques, l'intérêt pour les musées, et le budget alloué aux sorties culturelles. Le but est donc de modéliser l'effet additif de ces variables sur la probabilité de fréquenter le théâtre régulièrement ou non.

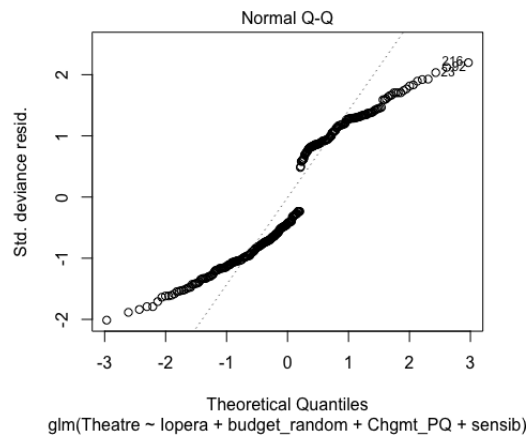
9. STEVEN, "Modèles additifs généralisés dans la modélisation de l'impact du kilométrage et de l'exposition au risque en assurance automobile".

10. STATISTICA, *Data Mining : Modèles Additifs Généralisés*.

### 5.2.1 Régression logistique

Afin de savoir s'il convient d'appliquer un modèle additif généralisé pour connaître la probabilité de fréquentation du théâtre, on utilise la fonction `gam()` du package *mgcv* qui établit une première relation par régression logistique afin d'analyser la distribution théorique des résidus, simulée par la régression. Le graphique ci-dessous nous montre la distribution théorique en abscisse et observée en ordonnée. Nous savons que les variables suivent une loi binomiale, dont les variables prédites par la modélisation permettent d'obtenir un vecteur des résidus pour chacune d'elle.

FIGURE 11 – Distribution théorique et observée des résidus du modèle logit sur le théâtre



L'objectif est donc de savoir s'il existe une non linéarité, en comparant le modèle logit avec le modèle linéaire généralisé.

### 5.2.2 Modèle additif généralisé

Nous avons montré dans la section précédente qu'il y avait une potentielle non-linéarité de la relation, au vu de la distribution des résidus. La loi de probabilité proposée est une loi binomiale avec une fonction de lien logistique<sup>11</sup>. Nous allons donc comparer le modèle linéaire proposé par le logit avec la fonction `glm()`<sup>12</sup>, avec le modèle additif généralisé obtenu par la fonction `gam()`. Si la comparaison montre un changement de déviance significative, on conclura à l'existence d'une non linéarité pour un gain en déviance significative au seuil de risque 0.1%<sup>13</sup>.

En tableau 8 ci-après, nous pouvons voir que l'edf, qui représente le nombre de degrés de liberté effectifs, est de 3.058. Plus ce nombre est élevé et plus la forme prise par la modélisation a une forme de courbe et non d'une droite. En effet, dans le cas d'une simple droite très lisse, un seul paramètre de pente serait estimé pour ajuster le lissage (edf=1), a contrario, pour les cas très complexes nous pourrions avoir autant de degrés de liberté qu'il y existe de points dans le tracé. Ce lissage est réalisé à partir des fonctions de base. L'objectif est donc d'optimiser le paramètre de lissage représentant la précision de la régression, noté  $\lambda$  et le nombre de fonctions de base noté  $k$  en respectant le critère de parcimonie.

11. C.Cans et C.Lavergne, "De la régression logistique vers un modèle additif généralisé : un exemple d'application, Revue de statistique appliquée, 1995

12. Confère 4.3.2

13. Ibidem

TABLE 8 – Estimation du modèle additif généralisé

Parametric coefficients	Coef	Err.Std	p-value
Constante	-2.638	0.753	0.000 ***
Changement de pratiques culturelles 1	1.096	0.617	0.076 .
Changement de pratiques culturelles 2	0.227	0.627	0.717
Intérêt pour l'opéra 1	0.951	0.294	0.001 **
Intérêt pour l'opéra 2	1.123	0.351	0.001 **
Intérêt pour les musées 1	-0.560	0.579	0.334
Intérêt pour les musées 2	-0.114	0.575	0.843
Sensibilisation par les parents	1.601	0.461	0.001 ***
Sensibilisation par l'école	0.983	0.473	0.038 *
Sensibilisation par les deux	1.700	0.433	0.000 ***
Approximate significance of smooth terms	edf	Chi.sq	p-value
Budget	3.058	14.49	0.005 **
Adjusted R <sup>2</sup>	0.203		
Deviance explained	19.3%		
Gam check	k'	k-index	p-value
s(budget)	9	1.04	0.74

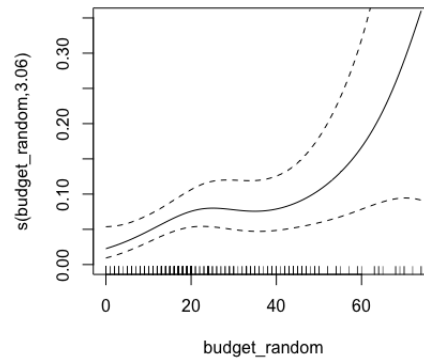
Ici, l'edf n'est pas proche de 1 alors nous savons que la forme linéaire n'est pas adaptée. La déviance expliquée est de 19.3%. L'estimation du gam nous donne la forme de la fonction  $s(\text{budget})$  appelée 'spline' ci-dessous, sur laquelle on peut observer l'effet de la variable budget sur l'échelle transformée du logit ; soit la probabilité de fréquentation du théâtre. On obtient donc l'effet de la variable budget par transformation du prédicteur sur la probabilité de fréquentation du théâtre. Nous pouvons voir que le prédicteur ne suit pas une forme linéaire. En effet, on observe des courbes avec une phase croissante jusqu'au point d'inflexion à 25 euros de budget, puis une courte phase décroissante jusqu'à 40 euros pour finir sur une phase ascendante. On constate en effet que l'intervalle de confiance augmente à partir de 50 euros, car les observations sont rares à ce niveau de budget. De même, nous pouvons voir que les barres en abscisses sont plus claires au fur et à mesure que le budget augmente (quand les barres sont en gras cela signifie qu'il y a plus d'observations). On note enfin, à partir de la table n°8, que la qualité d'ajustement du modèle est de 20% ce qui est relativement peu (quoique 3% de plus que le logit sur le théâtre), mais cela peut s'expliquer par le fait qu'un phénomène binaire est plus difficile à prédire.

En annexe n°7 on retrouve les différents graphiques des résidus pour le modèle GAM où l'on s'aperçoit avec l'histogramme que ceux-ci ne semblent pas suivre une loi normale. Ainsi, nous savons que la relation entre nos variables n'est pas linéaire. De plus, le QQ-plot nous indique que les résidus ne se trouvent pas au niveau des valeurs extrêmes mais plutôt au milieu.

Ensuite, nous pouvons voir grâce à la fonction '*gam.check*' que le nombre de degrés de liberté effectif (edf=3.058) est inférieur au nombre de fonctions de base (k=9) utilisées dans l'espace k. En effet, la p-value associée est supérieure à 0.05 ; le nombre de fonctions de base est donc adapté à la modélisation. De plus, les résidus sont distribués aléatoirement ce qui nous indique qu'il n'y a pas de pattern. Le paramètre de lissage est déterminé par le nombre de fonctions de base multiplié aux paramètres du modèle. Plus on a de fonctions

de base, mieux on peut approximer la courbe tout en faisant attention à ne pas trop complexifier le modèle plus que nécessaire. Ici, ce paramètre est égal à 1.04.

FIGURE 12 – Probabilités de la variable budget sur le théâtre



La figure n°12 confirme donc la non-linéarité de la relation entre le budget aléatoire et la fréquence du théâtre. Nous pouvons voir que la fréquentation du théâtre augmente avec le budget jusqu'à 20 euros. À partir de 20/25 euros, on peut voir que la courbe décroît jusqu'à son point d'inflexion à 35 euros environ où la courbe remonte progressivement de manière exponentielle. Ainsi, nous pouvons conclure de l'existence réelle d'un effet de seuil comme nous le soupçonnions au départ. Entre 20 et 35 euros de budget alloué aux sorties culturelles, la probabilité de fréquenter le théâtre diminue. Par précaution, nous avons également modélisé la fréquence du théâtre comme ayant pour seule variable explicative la fonction  $s(\text{budget})$ . Nous chercherons à comparer ces différents modèles ci-après grâce au test Anova.

Pareillement, il peut être intéressant de comparer sur quelles prédictions le modèle additif généralisé est meilleur que le logit, nous utilisons pour cela la fonction `hitmiss()` sous R qui nous informe des taux de prédictions correctes selon la modalité de  $Y$ . Le tableau n°9 ci-dessous regroupe l'information liée à la qualité de prévisions des différents modèles.

TABLE 9 – Taux de prédictions correctes selon le modèle estimé

	Modèle binaire	Modèle généralisé
Prédiction correcte	70.18%	71.99%
Prediction correcte pour $Y=0$	79.27%	76.17%
Prediction correcte pour $Y=1$	57.55%	66.19%

Nous observons ainsi qu'en moyenne, le modèle additif prédit près de 2 points de pourcentage mieux que le modèle logit. Cette différence est surtout marquée pour la prédiction de  $Y=1$  pour laquelle le modèle généralisé prédit correctement 66.19% tandis que le modèle binaire prédit correctement 57.55%, soit plus de 8 points d'écart. En revanche, la prédiction du fait de ne pas se rendre au théâtre est mieux appréhendée par le logit qui, 4 fois sur 5, prédit correctement cette probabilité. Le GAM, quant à lui, prédit justement un peu plus de 6 fois sur 8.

### 5.2.3 Conclusions sur les modélisations de $Y_1$

Ainsi, notre modèle semble adapté puisqu'il a révélé la non-linéarité de notre variable de budget randomisée. On utilise alors la fonction *Anova* pour statuer sur l'intérêt d'un modèle GAM par rapport à un simple modèle logit. Celui-ci se base sur l'égalité des moyennes et l'analyse de la variance pour comparer les modèles entre eux. Sa règle de décision est la suivante ;  $H_0$  il n'y a pas de différence significative entre les modèles testés et  $H_1$  il y a une différence significative. Il convient alors de choisir le modèle minimisant la variance résiduelle notée *Resid. Df.* Ainsi, en annexe n°8 on retrouve les résultats de ce test appliqué sur le modèle logit (n°1), le GAM composé de  $s(\text{budget})$  et des autres variables explicatives (n°2), et le modèle GAM avec seulement  $s(\text{budget})$  (n°3). Nous pouvons voir que l'hypothèse  $H_0$  est rejetée au seuil de 5% : il existe une différence significative entre les trois modèles. Le modèle minimisant la variance résiduelle est le modèle 2 *ie.* le modèle GAM composé de  $s(\text{budget\_random})$  et des autres variables explicatives avec une fonction linéaire par rapport à  $Y_1$ . Ajouter la fonction non linéaire a donc été fructueux par rapport au modèle logit, de même il apparaît que le modèle n°2 est plus pertinent que le 3, ce qui est logique puisqu'il contient plus d'informations apportées par les autres variables explicatives.

#### 1. Discrétisation de la variable 'budget\_random'

Pour pouvoir interpréter les coefficients de notre modèle, une des solutions possibles est de discrétiser la variable 'budget\_random' en différentes classes reprenant les seuils constatés et mis en avant à l'aide du GAM, puis de l'intégrer dans le modèle logit simple que nous avons construit et validé en partie précédente (cf. pages 23-24). Après avoir randomisé notre variable de manière à obtenir des valeurs quantitatives, nous reprenons sa manipulation pour la rendre qualitative et composée des différentes classes observées notamment sur la figure n°12 qui mettait en relation la variable du budget randomisé avec celle du théâtre. Nous tenions, avant de procéder aux manipulations, à rappeler que les conclusions que nous allons tirer du modèle logit où sera incluse la variable du budget discrétisée, dépendent fortement de la randomisation à laquelle nous avons procédé au début de l'analyse. Ainsi, nous n'affirmons pas que ces conclusions seront parfaitement justes au vu de notre échantillon de données, mais du moins qu'elles offrent une piste de réflexion sur l'influence du budget sur la pratique d'activité culturelle - et notamment la fréquentation des théâtres.

Discrétiser une variable quantitative, c'est la transformer en une variable qualitative ordinaire discrète par un découpage en intervalles à partir desquels on obtient différentes classes. Dans le cadre d'une étude supervisée, l'objectif de la discrétisation est de produire un ensemble d'intervalles où, pour chacun d'entre eux, certaines valeurs de la variable sont largement sur-représentées par rapport aux autres. Cependant, il est important d'avoir à l'esprit que ces transformations engendrent une perte d'information puisque certains paramètres ne seront plus calculables à partir d'une distribution discrète comme par exemple la moyenne, l'écart type... Effectivement, chaque classe définie par découpage regroupe des individus sous un même caractère, mais qui à l'origine se distinguaient par des valeurs différentes. C'est pourquoi le processus d'élaboration des classes est important, le but étant de synthétiser un volume important d'informations en limitant la perte liée à la discrétisation.<sup>14</sup>

Dans notre cas, nous n'effectuons pas ce découpage "à l'aveugle" puisque nous avons au préalable étudié notre variable à travers les statistiques descriptives, ainsi que le modèle additif généralisé qui nous a permis de vérifier notre hypothèse concernant l'existence de seuils sur la variable du budget alloué aux sorties culturelles.

14. EDUCATIM, *Transformation de variables qualitatives en variables quantitatives*.

Pour la cohérence des résultats, nous effectuons ce découpage sur la base libérée des 4 données influentes que nous avons constatées lors de la construction du modèle logit sur le théâtre,<sup>15</sup> puisque c'est dans ce modèle que nous incluons la variable discrétisée du budget.

Dans un premier temps, nous cherchons à discrétiser cette variable en utilisant la fonction 'mdlp()' contenue dans le package **discretization** sous R. Cette fonction suit une technique descendante (top down), et découpe de manière récursive l'espace des valeurs de  $X_i$  avec pour objectif de différencier au maximum les distributions des classes. Nous l'appliquons ainsi à notre jeu de données, et plus particulièrement à la variable "budget\_random", mais il apparaît qu'aucun intervalle n'est construit. Effectivement, grâce à la sortie R qui nous permet de voir les bornes de discrétisation, nous voyons que toutes les valeurs ont été mises dans une seule classe ("ALL"). Ainsi, le découpage est inefficace via cette fonction R puisque nous sommes intéressées en la création de 3 classes que nous avons pu souligner grâce à l'étude du modèle additif généralisé.

Par conséquent, dans un second temps nous utilisons une autre méthode de découpage : 'cut()', pour laquelle nous devons définir le nombre de classes à créer. Nous imposons ainsi 3 classes et cherchons à voir, par curiosité, si les intervalles correspondent à ceux constatés précédemment à savoir  $[0;25]$ ,  $[25;40]$  et  $[40;+\infty[$ .

TABLE 10 – Découpage de la variable "budget\_random" par la fonction 'cut' avec 3 classes imposées

Bornes des classes créées	] -0.074 ; 24.7]	] 24.7 ; 49.3]	] 49.3 ; 74.1]
Nombre d'observations	216	90	26

D'après la table 10, il apparaît que les classes ne sont pas tout à fait comme celles que nous avons constatées visuellement. Le découpage n'est pas parfait puisque l'on voit déjà que la première classe débute en -0.074, alors que les valeurs de la variable initiale s'étendent de 0 à 74. En revanche, la borne 25 est conservée mais l'autre borne est autour de 50 tandis que nous l'avions définie à 10€ de moins. Nous procédons donc au découpage 'final', puisque ce dernier nous permettait uniquement à voir quelles classes la fonction 'cut' créait si celles-ci n'étaient pas définies. À présent, nous imposons le découpage mis en avant avec le modèle GAM, en procédant à la discrétisation du budget comme suit :

TABLE 11 – Découpage final de la variable "budget\_random"

Bornes des classes créées	[0 ; 25]	] 25 ; 40]	] 40 ; 74]
Nombre d'observations	219	51	62

Ainsi, le budget est maintenant composé de 3 modalités correspondant aux 3 classes visibles en tableau 11 ci-dessus. Comparé au découpage précédent, on voit que la répartition dans les classes 2 et 3 est meilleure étant donné le plus grand nombre d'observations en modalité n°3 (pour un budget de 41€ à 74€). Nous pouvons alors enfin inclure cette variable dans le logit, calculer les odd ratios et interpréter les résultats.

15. La base contient ainsi 332 observations, au lieu de 336.

2. Interprétations des résultats

TABLE 12 – Odd ratios des coefficients du modèle logit sur le théâtre

Variables	$P[Y_1=1]$	p-value
Budget culturel entre 26 et 40€	1.160	0.677
Changement de pratiques culturelles 2	1.266	0.690
Budget culturel entre 41 et 74€	2.156	0.021 *
Intérêt pour l'opéra 1	2.748	0.000 ***
Sensibilisation par des activités extra-scolaires	2.881	0.023 *
Changement de pratiques culturelles 1	3.046	0.055 .
Intérêt pour l'opéra 2	3.288	0.000 ***
Sensibilisation par les parents	5.723	0.000 ***
Sensibilisation par les deux	6.248	0.000 ***

Les rapports de chances (odd ratios) disponibles en table n°12 et triés par ordre croissant, nous informent que les variables significatives à 5% ayant l'effet le plus grand sur la probabilité qu'un étudiant se rende au théâtre, sont le fait d'avoir été sensibilisé par les parents et/ou des cours extra-scolaires, ainsi que le fait de s'intéresser à l'opéra. Nous pouvons désormais interpréter les odd ratios des variables étant significatives dans l'explication de la fréquentation de théâtres.

Ainsi, les étudiants avec un budget allant de 41 à 74 euros, ont eue 2.156 fois plus de chance de se rendre au théâtre, par rapport à des individus avec un budget de moins de 26 euros (classe de référence de la variable). L'effet de seuil semble donc avoir été évincé à nouveau puisque la probabilité de fréquentation du théâtre augmente avec le budget. Effectivement, bien que non significative nous voyons que le rapport de chance de la variable budget compris entre 26 et 40€, est plus faible que celui compris entre 41 et 74€ que nous venons d'interpréter. Cela peut s'expliquer par la pente négative identifiée entre 26 et 40 euros sur la figure 12. Concernant l'intérêt pour l'opéra, les modalités sont toutes deux largement significatives : les étudiants moyennement intéressés *i.e.* ayant répondu 'pourquoi pas', ont 2.748 fois plus de chance d'aller au théâtre au moins 1 fois par an, que les étudiants pas intéressés du tout par l'opéra. Idem pour les étudiants très intéressés qui ont cette fois 3.288 fois plus de chance par rapport aux non intéressés.

De plus, la volonté de consacrer plus de temps aux pratiques culturelles augmente la probabilité d'aller au théâtre au moins une à deux fois par an de 3.046 par rapport aux étudiants ne souhaitant rien changer de leurs pratiques existantes (au seuil de risque de 10%). En revanche, la volonté de diversifier ses pratiques culturelles (modalité n°2), ne sort pas significative du modèle logit donc n'a pas d'influence sur le fait de se rendre au théâtre au moins une fois l'an. Pour finir, les étudiants ayant été sensibilisés à la culture par les activités extra-scolaires ont 2.881 fois plus de chances de fréquenter les théâtres, par rapport à des personnes n'ayant pas été sensibilisées. Les personnes sensibilisées par leurs parents ont eue 5.723 fois plus de chance et les personnes ayant été sensibilisées par les deux à la fois ont alors 6.248 fois plus de chance que des personnes n'ayant jamais été sensibilisées. Aussi, les habitudes culturelles passant par l'éducation des enfants par les parents ont plus d'influence sur les futures pratiques culturelles, que seulement des cours extra-scolaires de musique, peinture, dessin, danse etc. En revanche, cet effet est considérable lorsque les étudiants ont été sensibilisés à la Culture via ces 2 possibilités puisqu'ils ont alors 6.248 fois plus de chances de se rendre au



théâtre comparé à un étudiant n'ayant pas du tout été sensibilisé.

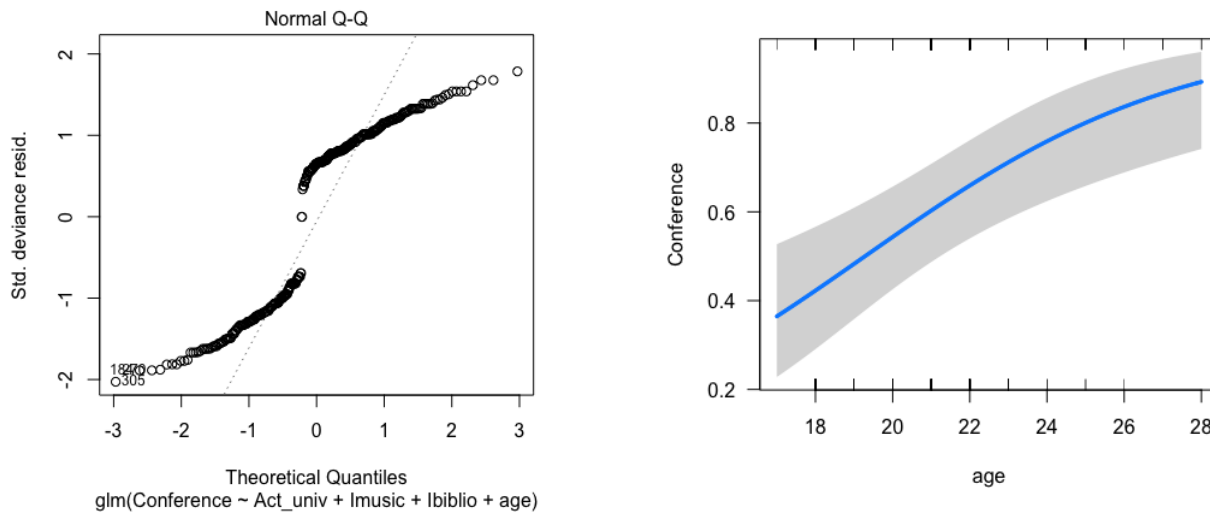
Ces constats confirment nos hypothèses émises en partie économique où nous cherchions notamment à savoir si les pratiques culturelles étaient le résultat d'habitudes transmises par l'éducation. De plus, l'intérêt pour l'opéra montre que les pratiques culturelles sont diverses et variées mais restent liées entre elles. En outre, quand un individu s'intéresse à une pratique culturelle, sa curiosité envers les autres formes d'art augmente et celui-ci sera donc plus enclin à pratiquer une autre activité culturelle.

### 5.3 Application sur la fréquentation des conférences

#### 5.3.1 Régression logistique

Nous allons réaliser la même analyse sur le fait de se rendre à des conférences, sachant que nous avons déjà estimé un modèle logit binaire sur cette variable. Nous cherchons donc à savoir si un modèle GAM, c'est à dire avec une fonction non linéaire sur la variable de l'âge, est plus pertinent et plus précis qu'un simple modèle logit. Nous commençons donc par regarder l'aspect visuel des résidus du modèle binaire pour avoir une idée de la loi suivie, ainsi que la nature de la relation reliant  $Y_2$  à l'âge.

FIGURE 13 – Diagnostiques du modèle logit sur les conférences : résidus et relation âge/ $Y_2$



Nous voyons alors que la distribution théorique des résidus du modèle logit s'apparente effectivement à la fonction de répartition de la loi binomiale comme pour le modèle du théâtre. Les résidus ne semblent en effet pas linéaires, puisque les points ne sont pas alignés sur la droite QQ. De plus, il apparaît sur le graphique de droite que la relation entre l'âge et les conférences n'est pas tout à fait linéaire. On voit que de 17 à 23 ans, plus l'âge du répondant augmente plus la probabilité qu'il se rende à des conférences augmente à son tour - linéairement à l'augmentation de l'âge. En revanche, pour les étudiants âgés de 23 à 28 ans on constate qu'à mesure que leur âge augmente la probabilité d'aller aux conférences augmente de moins en moins rapidement. On observe alors comme un effet plafond où la probabilité de  $Y_2$  semble stagner autour de 0.9.

Nous allons vérifier cette linéarité ou non linéarité de la relation qui lie l'âge aux conférences, par

la modélisation d'un modèle GAM composé des 3 variables explicatives linéaires, avec l'âge sous fonction de lissage non-paramétrique. Nous verrons alors si le fait d'ajouter cette fonction non spécifique permet d'améliorer significativement l'explication du fait de se rendre à des conférences.

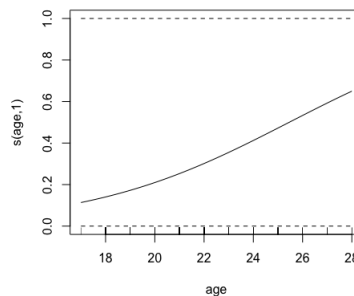
### 5.3.2 Modèle additif généralisé

TABLE 13 – Estimation du modèle additif généralisé sur les conférences

Parametric coefficients	Coef	Err.Std	p-value
Constante	-1.063	0.347	0.002 **
Participation aux activités des établissements 1	0.740	0.251	0.003 **
Participation aux activités des établissements 2	-49.999	4.75e+07	1.000
Intérêt pour la musique classique 1	0.871	0.292	0.003 **
Intérêt pour la musique classique 2	0.774	0.308	0.012 *
Intérêt pour les bibliothèques 1	0.630	0.323	0.051 .
Intérêt pour les bibliothèques 2	-0.786	0.337	0.020 *
Approximate significance of smooth terms	edf	Chi.sq	p-value
Âge	1	1	0.000 ***
Adjusted R <sup>2</sup>	0.118		
Deviance explained	11%		
Gam check	k'	k-index	p-value
s(age)	9	0.91	0.05 *

Nous constatons d'emblée que le degré de liberté du terme de lissage est de 1. Comme nous l'avons expliqué précédemment, cela signifie que la modélisation est linéaire. Nous voyons par ailleurs que ce paramètre est significatif dans l'estimation du GAM et que ses résidus sont distribués aléatoirement puisque au seuil de risque de 10% puisque la p-value associée au coefficient  $s(\text{age})$  est de 0.05. La linéarité de la relation liant l'âge aux conférences démontrée par la valeur du degré de liberté, peut être confirmée par la figure n°14 ci-dessous qui montre effectivement une linéarité de la relation. On observe une très légère convexité de la courbe mais l'on voit bien que l'âge affecte la probabilité de se rendre à des conférences de manière positive et linéaire ; quand un étudiant est âgé de 20 ans, la probabilité qu'il aille à des conférences est de 0.2 en revanche pour un étudiant de 27 ans elle sera 3 fois plus élevée c'est à dire de 0.6.

FIGURE 14 – Relation entre l'âge et les conférences



L'estimation d'un modèle additif généralisé ne permet donc pas d'améliorer la qualité de la prévision de  $Y_2$  mais complexifie l'analyse en incluant une fonction non linéaire : nous vérifierons cette hypothèse en appliquant le test Anova qui confronte la précision des différents modèles et permet de voir lequel est statistiquement meilleur. Nous ne modélisons pas d'autres modèles généralisés composés par exemple seulement de la variable  $s(\text{age})$  comme nous avons pu le faire pour les sorties au théâtre, puisque nous avons vu qu'a priori la relation liant l'âge aux conférences est linéaire et ne nécessite donc pas de modèle généralisé.

Disponible en annexe n°9, le test Anova révèle au seuil de risque de 5% qu'il y a une différence significative entre le modèle logit et le modèle linéaire généralisé. Il convient donc de choisir le modèle ayant la variance la plus faible : nous voyons alors que cette dernière est identique entre les 2 modèles, de même pour l'argument "*Resid. Dev*". Dans un souci de parcimonie et de simplicité d'interprétation, nous décidons de garder le modèle logit plutôt que le modèle additif généralisé. Effectivement, le modèle binaire étant constitué uniquement de fonctions linéaires est plus simple à interpréter. Nous allons, dans une dernière sous partie, interpréter les résultats de la modélisation de  $Y_2$  par le logit que nous avons pu estimer en partie 4.3.1.

Nous rappelons que le modèle sélectionné prédit correctement 77.04% du fait de se rendre à des conférences, et 50% du fait de ne pas s'y rendre. Il est donc meilleur dans la prédiction de la probabilité  $P[Y_2=1]$ .

### 5.3.3 Interprétation des résultats

S'agissant d'un modèle binaire, les résultats issus de la sortie informatique du modèle logit disponible en table n°6 ne sont pas interprétables directement. Nous commençons donc par calculer les odd ratios des variables qualitatives et l'effet marginal de la variable quantitative, de manière à conclure sur l'explication du fait de se rendre à des conférences - reflétant la curiosité des étudiants nantais vis-à-vis d'une des nombreuses formes de la Culture. Par ailleurs, nous pouvons dès à présent regarder d'après la table n°6 d'estimation du modèle binaire, quelles variables sont significatives ainsi que leur signe, reflétant l'impact positif ou négatif sur la probabilité de se rendre à des conférences. Aussi, nous constatons que chaque modalité des facteurs explicatifs a un effet positif sur la probabilité d'aller aux conférences, excepté ceux qui ne savent pas s'ils participent aux activités proposées par leur établissement - modalité cependant non significative. Cela est largement compréhensible étant donné la faible représentativité de l'échantillon dans cette modalité. Elle ne regroupe effectivement que 2 étudiants qui ne savaient pas s'ils participaient aux activités de leur pôle ou BDA. Nous ne prendrons donc pas en compte la modalité de cette variable dans les interprétations du modèle.

TABLE 14 – Effets marginaux et odd ratios des coefficients du modèle logit

	Variables	$P[Y_2=1]$	p-value
<b>Odd ratios</b>	Intérêt pour les bibliothèques 1	1.877	0.051 .
	Participation aux activités des établissements 1	2.096	0.003 **
	Intérêt pour la musique classique 2	2.168	0.012 *
	Intérêt pour les bibliothèques 2	2.194	0.020 *
	Intérêt pour la musique classique 1	2.388	0.003 **
<b>Effet marginal</b>	Âge du répondant	0.051	0.000 ***

Nous voyons en table n°14 les coefficients transformés des variables que nous pouvons désormais interpréter. Ceux-ci sont classés par ordre croissant de leur effet sur  $P[Y_2=1]$ . Les odd ratios donnent la probabilité

d'aller à des conférences pour un étudiant par rapport à la modalité n°0 des variables concernées : ne pas être intéressé par les activités pour les variables d'**intérêt**, et ne pas participer aux activités des établissements.

Ainsi, un étudiant étant intéressé par des sorties à la bibliothèque ou à la médiathèque sans y être réellement intéressé (modalité "pourquoi pas" en réponse à la question sur l'intérêt porté à cette activité) aura 1.877 fois plus de chance de se rendre à des conférences par rapport à un étudiant qui n'aimerait pas du tout s'y rendre. Pour les individus étant fortement intéressés par des sorties à la bibliothèque, leur chance de se rendre à des conférences par rapport à des étudiants qui ne veulent pas aller dans les bibliothèques, est alors 2.194 plus grande. Bien que cela tombe sous le sens, il est intéressant de noter les relations entre les activités culturelles ; plus une personne est intéressée par certaines activités, plus sa chance de se rendre à d'autres (des conférences ici) augmente.

Nous notons aussi qu'un étudiant participant aux activités culturelles proposées par son établissement a 2.096 fois plus de chance de se rendre à des conférences par rapport à un étudiant n'y participant pas. Cela semble logique aussi, mais confirme la pertinence de cette variable qui permet de cerner les pratiques des étudiants nantais lorsque l'information des événements culturels vient jusqu'à eux. Concernant l'intérêt pour la musique classique, nous voyons cette fois qu'un étudiant étant fortement intéressé a 2.168 fois plus de chance de se rendre à des conférences que quelqu'un n'étant pas du tout intéressé, tandis que ce chiffre est de 2.388 pour les individus étant ouverts à la musique classique sans vouloir beaucoup en écouter. L'effet est cette fois décroissant sur la probabilité de  $Y_2$  lorsque l'intérêt pour la musique classique augmente. Enfin, comme nous l'avions remarqué avec les diagrammes en barres dans les statistiques descriptives de notre base, au fur et à mesure que l'âge du répondant augmente, sa probabilité de se rendre à des conférences augmente elle aussi. Concrètement, 1 an de plus par rapport à l'âge moyen des étudiants augmente de 0.051 la probabilité d'aller aux conférences.

Nous avons ainsi vu quels facteurs influencent la probabilité de se rendre à des conférences et comment ils l'influençaient. Assemblée à l'analyse du fait d'aller au théâtre, nous avons pu étudier, dans ces 2 parties, les déterminants des activités culturelles représentant une sortie et qui se pratiquent occasionnellement. Cependant, nous savons que la plupart des étudiants, ayant un petit budget, préfèrent aux sorties culturelles les pratiques quotidiennes. Nous allons alors, dans une nouvelle partie, étudier les déterminants des activités culturelles au quotidien.

## 6 Modèle probit multivarié

### 6.1 Fonctionnement théorique d'un modèle probit multivarié

Le modèle probit multivarié comme extension au modèle probit simple est utilisé pour analyser les choix multiples.<sup>16</sup> On cherche à expliquer une variable associée à des choix binaires tels que 'OUI/NON'. Comme pour le biprobit, les choix ne sont pas indépendants et exclusifs mais au contraire, sont corrélés entre eux<sup>17</sup>.

On considère alors dans un probit multivarié un nombre M d'équations selon le nombre de variables que l'on estime liées, tel que<sup>18</sup> :

$$y_{im}^* = \beta'_m X_{im} + \epsilon_{im}, m = 1, \dots, M$$

Où  $y_{im} = 1$  si  $y_{im}^* > 0$  ou 0 dans le cas contraire.

$\epsilon_{im}, m = 1, \dots, M$  représentent les termes d'erreurs, dont la moyenne est de zéro et qui par hypothèse suivent une loi normale multidimensionnelle. De plus, la matrice de variance-covariance V est composée de valeurs '1' sur la diagonale principale et des corrélations  $\rho_{jk} = \rho_{kj}$  en dehors.

Dans le cas d'un modèle triprobit on a alors M=3, ce qui nous donne 8 combinaisons de probabilités jointes possibles de succès (1) et d'échec (0). On présume donc l'existence de trois variables latentes tel que<sup>19</sup> :

$$y^*_{i1} = X_{i1}\beta_1 + \epsilon_{i1}$$

$$y^*_{i2} = X_{i2}\beta_2 + \epsilon_{i2}$$

$$y^*_{i3} = X_{i3}\beta_3 + \epsilon_{i3}$$

où  $X_{ij}$  sont les vecteurs des prédicteurs,  $\epsilon_{ij}$  est le terme aléatoire et  $\beta_j$  est le vecteur des coefficients estimés pour j=1,2,3.

$$X_{ij} = \begin{bmatrix} x_{i1j} & \dots & x_{iKj} \end{bmatrix}, \text{ vecteur de dimension } (1 \times K) \text{ et } \beta_j = \begin{bmatrix} \beta_{j1} \\ \vdots \\ \beta_{jK} \end{bmatrix}, \text{ de dimension } (K \times 1).$$

$$y_{i1} = \begin{cases} 1 & \text{si } y^*_{i1} > 0 \\ 0 & \text{si } y^*_{i1} \leq 0 \end{cases}$$

$$y_{i2} = \begin{cases} 1 & \text{si } y^*_{i1} > 0, y^*_{i2} > 0 \\ 0 & \text{si } y^*_{i1} > 0, y^*_{i2} \leq 0 \\ \text{non observé} & \text{si } y^*_{i1} \leq 0 \end{cases}$$

$$y_{i3} = \begin{cases} 1 & \text{si } y^*_{i1} > 0, y^*_{i2} > 0, y^*_{i3} > 0 \\ 0 & \text{si } y^*_{i1} > 0, y^*_{i2} > 0, y^*_{i3} \leq 0 \\ \text{non observé} & \text{si } y^*_{i1} \leq 0, \text{ ou } y^*_{i2} \leq 0 \end{cases}$$

16. AURIER et MEJÍA, "Les modèles Logit et Probit multivariés pour la modélisation des achats simultanés".

17. N'GUESSAN, "Analyse des déterminants de l'intensité de la recherche d'emploi en Côte d'Ivoire".

18. CAPPELLARI et JENKINS, "Multivariate probit regression using simulated maximum likelihood".

19. CARREÓN et L.GARCÍA, "Trivariate Probit with Double Sample Selection : Theory and Application".

## 6.2 Application sur la lecture, la radio et les jeux vidéo

Afin de compléter notre analyse, nous nous sommes demandé quels étaient les déterminants des pratiques culturelles quotidiennes telles que le fait de lire, d'écouter la radio ou de jouer aux jeux vidéo. Nous connaissions à quelles fréquences les étudiants nantais pratiquaient ces activités mais nous n'avions à jour fait aucun test statistique pour en connaître les raisons. Tout comme le fait de fréquenter les théâtres peut-être lié au fait de se rendre à des conférences, la lecture, la radio et les jeux-vidéo sont des activités qui peuvent se substituer. En effet, on sait instinctivement que ces pratiques sont le résultat d'habitudes que l'on aurait principalement chez soi et qui seraient donc liées. En outre, nous userons d'un modèle probit multivarié afin d'expliquer la probabilité de pratiquer une de ces activités par rapport aux deux autres. Il s'agira alors d'un modèle à trois variables à expliquer ; soit un triprobit.

Contrairement aux pièces de théâtre et aux conférences, lire, écouter la radio ou encore jouer à des jeux vidéo sont des activités qui se pratiquent plutôt quotidiennement. Ainsi, la répartition des effectifs au sein des modalités diverge entre ces deux catégories d'activités et nous avons décidé afin d'obtenir une equirépartition, de séparer les individus pratiquant l'activité au moins une fois par semaine de ceux la pratiquant moins souvent. Pour résumer : les modalités 0, 1, 2, 3 *ie.* les étudiants pratiquant l'activité moins d'une fois par mois voire jamais, prennent la modalité 0. Les étudiants pratiquant l'activité au moins une fois par semaine prennent la modalité 1. De plus, les étudiants ayant répondu "ne sait pas" ont été affectés à la modalité 0 puisque l'on considère que si la pratique était récurrente le constat leur serait plus évident.

Ainsi, nous obtenons les répartitions suivantes pour chaque modalité où '*Oui*' : activité pratiquée une fois ou plus par semaine, et '*Non*' dans le cas contraire *ie.* moins d'une fois par semaine.

TABLE 15 – Tableau de contingence

		Jeux-Vidéo	
		Non	Oui
Lecture	Radio		
Non	Non	65 (19%)	22 (6%)
	Oui	40 (6%)	21 (6%)
Oui	Non	42 (25%)	31 (9%)
	Oui	83 (25%)	32 (9%)

Nous pouvons voir que les 8 modalités possibles semblent globalement bien réparties. La répartition parfaite ici encore, voudrait que chaque modalité croisée contienne 12.5% des observations. Nous constatons que les répartitions effectives s'étendent de 6% pour les individus ne lisant pas mais écoutant la radio (qu'ils jouent à des jeux vidéo ou non), à 25% pour les étudiants lisant, ne jouant pas aux jeux vidéo - qu'ils écoutent la radio ou non. On remarque par ailleurs, que sur notre échantillon, la majorité des répondants ne jouent pas aux jeux vidéo. De la même manière que pour le modèle biprobit, le test du Khi-2 utilisé sur notre tableau de contingence à trois variables nous indique si la distribution des variables dans l'échantillon est due au hasard. Sachant que la p-value est inférieure à 0.05 (p-value=0.01), l'hypothèse  $H_0$  est rejetée, il existe donc bien un lien entre nos 3 variables.

Pour modéliser ce lien entre nos variables nous utiliserons donc un triprobit. Nous supposons pour commencer que le terme d'erreur  $\rho$  de la matrice de variance-covariance suit une loi normale. Dans le cas

contraire nous ne pourrions pas appliquer ce modèle car cela voudrait dire qu'il n'existe pas de phénomène conjoint entre nos trois variables. Il faudrait alors appliquer des logits séparés. Dans le but de modéliser ces 3 phénomènes, et comme pour le biprobit, nous avons utilisé les méthodes stepwise, backward, et forward pour déterminer les variables pertinentes à inclure dans chacune des équations. Nous avons obtenu les résultats suivants :

TABLE 16 – Sélection de variables pour les  $Y_i$ 

$Y_1$ : La lecture	$Y_2$ : La radio	$Y_3$ : Les jeux-vidéo
Intérêt pour les bibliothèques	Intérêt pour le théâtre	Le genre
Intérêt pour la musique classique	Participation aux activités culturelles universitaires	L'âge
Intérêt pour la danse	Budget alloué aux sorties culturelles (aléatoire)	
L'âge		

Nous pouvons voir qu'hormis 'l'âge', on ne retrouve pas les mêmes variables pour expliquer la lecture, l'écoute de la radio ou le fait de jouer aux jeux-vidéo. On retrouve le genre dans les variables explicatives des jeux-vidéo, ce qui n'est pas étonnant étant une activité bien plus répandue chez nos homologues masculins. Concernant la lecture, on retrouve logiquement l'intérêt pour les bibliothèques. Le regroupement des modalités ainsi que les méthodes de sélection ont été réalisés sous *R*. Pour modéliser le modèle triprobit nous avons ensuite utilisé *Stata*.

TABLE 17 – Estimation du modèle triprobit

	Variables	Coef	Err.Std	p-value
<b>Lecture</b>	Intérêt pour les bibliothèques	0.376	0.099	0.000 ***
	Intérêt pour la musique	0.274	0.093	0.003 ***
	Intérêt pour la danse	0.192	0.091	0.036 **
	Age	0.069	0.036	0.053.
	Constante	-2.237	0.783	0.004 ***
<b>Radio</b>	Intérêt pour le théâtre	0.231	0.097	0.017 **
	Participation aux activités culturelles universitaires	-0.282	0.140	0.044 **
	Constante	-2.266	0.169	0.116
<b>Jeux vidéo</b>	Genre	0.946	0.155	0.000 ***
	Age	-0.101	0.038	0.009 ***
	Constante	1.303	0.805	0.105
<b>Corrélations</b>				
	$\rho^{12}$	0.249	0.089	0.004 **
	$\rho^{13}$	0.131	0.089	0.140
	$\rho^{23}$	-0.072	0.091	0.432
	Test de Wald	86.01		0.000
	Test de $\rho=0$	11.594		0.009
	Log likelihood = -623.173			

Note : 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Nous pouvons voir que le modèle triprobit est pertinent puisque cette fois-ci la p-value associée au test du ratio de vraisemblance est inférieure à 0.01. En outre, l'hypothèse  $H_0$  ne peut-être acceptée au seuil de risque de 1%, il est pertinent de modéliser les trois phénomènes ensemble car ils sont liés ; les étudiants nantais peuvent à la fois pratiquer la lecture, l'écoute de la radio mais également jouer aux jeux vidéo. La qualité d'ajustement de notre modèle est de 0.0092 soit 0.92% ( $R^2 = 1 - \frac{\loglikelihood}{restr.\loglikelihood} = 1 - \frac{-623.17292}{-628.96972}$ ).

Afin d'avoir une idée précise de l'impact de ces variables sur nos activités quotidiennes, il advient de calculer les effets marginaux. Nous obtenons ainsi leur impact sur la probabilité de passer d'une fréquence de pratique de moins d'une fois par mois à plus d'une fois par semaine sur les 8 probabilités conjointes. Or, il n'est pas possible de calculer ces effets avec la fonction *'mfx'* utilisée pour le biprobit. Ainsi, le logiciel R propose une fonction *mvProbit* pour estimer les probits multivariés. Seulement, cette fonction ne permet pas de modéliser les phénomènes à expliquer avec des variables explicatives différentes, ce qui est le cas dans notre modèle puisque nous n'avons sélectionné aucune variable explicative en commun. Une des solutions est donc d'assembler toutes les variables  $X_i$  pour expliquer nos phénomènes.

Après de nombreuses tentatives vaines sous R comme sous Stata pour obtenir les effets marginaux des variables, nous nous contenterons finalement uniquement d'interpréter la nature du lien entre les équations du modèle, ainsi que les signes des coefficients estimés. Nous regardons pour cela les coefficients des  $\rho$  qui représentent ces liens entre les équations. Ainsi, il apparaît que la pratique de la lecture favorise la pratique de la radio puisque  $\rho^{12}$  est positif. De même, si un étudiant nantais lit régulièrement - soit au moins 1 fois par semaine, sa chance de faire des jeux vidéo est plus grande et vice versa. Enfin, un étudiant écoutant au moins une fois par semaine la radio a moins de chance de faire des jeux vidéo ( $\rho^{23} = -0.072$ ) qu'un étudiant écoutant la radio à une fréquence plus réduite.

Aussi, bien que nous n'ayons pu trouver les effets marginaux de chaque variable sur les différents phénomènes à expliquer, nous pouvons d'ores et déjà regarder le signe des coefficients, ainsi que la significativité des variables. De ce fait, on observe que l'intérêt que les étudiants portent à la pratique d'activités culturelles telles que les bibliothèques, la musique et la danse, influencent positivement la chance de lire au moins une fois par semaine. Si nous considérons le seuil de risque de 10%, on voit qu'au fur et à mesure que l'âge augmente, la probabilité de lire régulièrement augmente à son tour. Concernant l'écoute de la radio, cette fois c'est l'intérêt porté au théâtre qui augmente les chances d'écoute. En revanche, on voit que si le répondant participe aux activités culturelles proposées par son établissement secondaire, alors la probabilité qu'il écoute la radio diminue : cela peut s'expliquer par une réduction du temps disponible pour écouter la radio si l'étudiant participe aux activités culturelles. Finalement, il apparaît que les garçons ont plus de chance de faire des jeux vidéo que les filles - constat que nous avons déjà fait précédemment, et que les jeunes étudiants ont aussi une probabilité plus forte d'y jouer. Ainsi, même si nous n'avons pu quantifier les effets de chaque variable sur les trois  $Y_i$ , nous avons pu constater les variables ayant un effet à la hausse ou à la baisse sur les chances de pratiquer les 3 activités culturelles quotidiennes.



## 7 Conclusion

Cette étude complémentaire nous aura permis de tirer de nouvelles conclusions grâce aux modélisations réalisées sur notre enquête concernant les pratiques culturelles des étudiants nantais. L'objectif initial de ce dossier était de modéliser conjointement la fréquentation du théâtre et des conférences, car nous supposions que l'intérêt porté à une des activités influençait la probabilité de pratiquer l'autre. Seulement, les termes d'erreurs du biprobit n'étant pas corrélés, il n'était donc pas pertinent de modéliser ces deux phénomènes ensemble. Nous avons donc appliqué séparément des modèles logit pour cerner quels étaient les déterminants de la fréquentation du théâtre et des conférences.

Concernant la probabilité de fréquentation du théâtre, nous soupçonnions la variable 'budget\_random' sélectionnée, d'avoir une relation non-linéaire avec la fréquence du théâtre. En effet, le modèle multinomial ordonné réalisé dans le dossier précédent avait montré qu'il existait des effets de seuil sur la variable de budget initial (en classes). Ainsi, nous avons complété le logit avec un modèle GAM qui a démontré qu'il existait bien une relation non linéaire, car de 27 à 40 euros la probabilité de se rendre au théâtre au moins une fois par an baisse avec le budget. Afin de pouvoir quantifier cet effet, nous avons discrétisé la variable après l'avoir randomisée, afin de l'inclure dans la modélisation logit initiale. Le modèle n'a pas montré de relation négative entre la tranche [27-40 euros] et la probabilité d'aller au théâtre, l'effet de seuil est peut-être trop léger pour se manifester dans ce modèle.

Les résultats montrent que plus l'intérêt pour l'opéra et la volonté de consacrer plus de temps aux pratiques culturelles augmentent, plus la probabilité de se rendre au théâtre une fois par an augmente à son tour, par rapport à des personnes non intéressées. Ces résultats sont cohérents avec nos hypothèses de départ puisque plus un individu s'intéresse à la culture sous n'importe quelle forme, plus il a de chance de faire des sorties culturelles. Nous avons également pu confirmer l'impact de la sensibilisation à la culture évoquée notamment dans les travaux du sociologue Pierre Bourdieu sur la probabilité de fréquenter le théâtre. En effet, le fait d'avoir été sensibilisé à la fois par les parents et des activités extra-scolaires augmente de 6.248 fois la probabilité de fréquenter le théâtre au moins une fois par an, ce, par rapport à un étudiant n'ayant pas été sensibilisé ! Concernant la probabilité de fréquentation des conférences, il s'est avéré que la variable d'âge n'avait pas de relation non-linéaire avec cette probabilité et donc que la modélisation logit suffisait. Ici, l'intérêt pour les bibliothèques comme pour la musique classique augmentait la probabilité de se rendre à des conférences au moins une fois par an, par rapport au fait de ne pas être intéressé. De plus, par rapport à l'âge moyen des étudiants qui est de 21 ans, un an d'âge supplémentaire augmente la probabilité de 0.051 de se rendre à des conférences. Ce qui semble également logique puisqu'intuitivement on se doute que l'intérêt pour les conférences dépend également du développement intellectuel d'un individu.

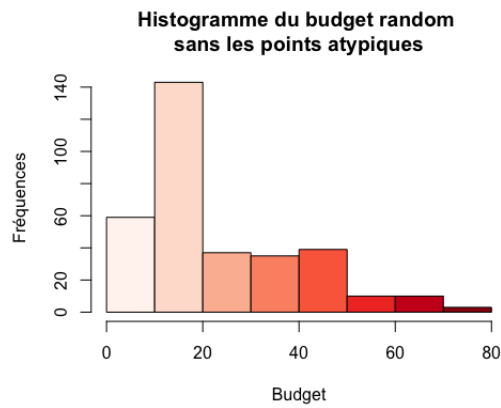
Pour finir, le modèle triprobit a montré que les activités pratiquées quotidiennement : la lecture, l'écoute de la radio et le fait de jouer aux jeux vidéo peuvent être modélisées ensemble puisque ces variables sont liées et peuvent donc s'influencer mutuellement. Nous avons alors constaté que les différentes variables significatives avaient principalement un effet positif sur la chance de pratiquer la lecture, les jeux vidéo ou l'écoute de la radio au moins une fois par semaine. Seuls la participation aux activités culturelles universitaires et l'âge du répondant diminuaient significativement respectivement les probabilités d'écoute de la radio et de jeux vidéo.

Notre objectif en reprenant cette enquête sur les pratiques culturelles des étudiants nantais commencée l'année dernière, était de compléter l'analyse pour mieux cerner le sujet. Cet objectif a donc été atteint puisque l'an passé nous nous étions principalement concentrées sur l'explication binaire des fréquentations

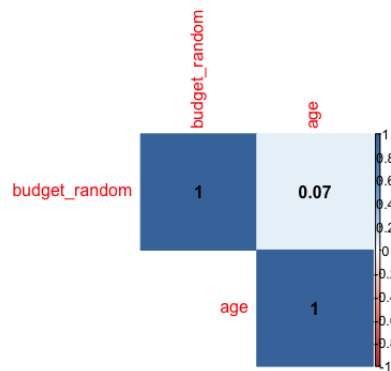
des théâtres ainsi que la pratique générale d'activités culturelles où nous avons attribué un score s'y référant à chaque étudiant. Cette année, grâce aux nouvelles techniques d'analyse étudiées, nous avons pu modéliser conjointement certains phénomènes et considérer la possible non-linéarité des relations entre certaines variables.

## 8 Annexes

**Annexe n°1** : Histogramme de distribution du budget randomisé sans ses valeurs atypiques.



**Annexe n°2** : Matrice de corrélation sur la base nettoyée.



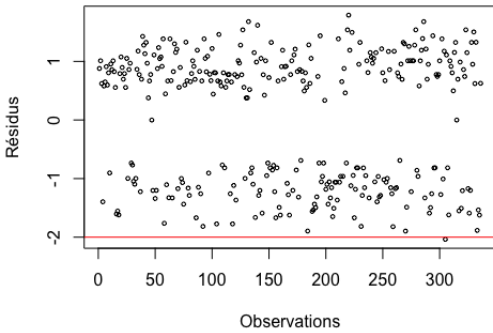
**Annexe n°3** : Tableau croisé des fréquences de sorties à des conférences avec le budget, l'âge et le sexe.

Fréquences	0	1	2	3	4	5
Âge moyen	20.49	21.36	21.64	22.67	22.33	23
Budget moyen	21.27	24.12	26.77	22.5	25.33	42
Femmes	100	79	41	3	3	1
Hommes	40	51	15	3	0	0

**Annexe n°4** : Tableau croisé des fréquences de sorties au théâtre avec le budget, l'âge et le sexe.

Fréquences	0	1	2	3	4	5
Âge moyen	21.16	20.9	21.06	23.67	20	23
Budget moyen	21.06	26.01	29.06	33.67	18	42
Femmes	124	87	12	2	1	1
Hommes	69	34	5	1	0	0

## Annexe n°5 : Démarche du modèle logit sur les conférences.

Modèle aux erreurs hétéroscédastiques avec toutes les variables en coupable	Modèle aux erreurs hétéroscédastiques avec la seule variable significative																																																																																																																																																	
<p>Coefficients (binomial model with logit link):</p> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>(Intercept)</td><td>-2.689e+00</td><td>5.301e+00</td><td>-0.507</td><td>0.612</td></tr><tr><td>Act_univ1</td><td>1.875e-01</td><td>3.640e-01</td><td>0.515</td><td>0.606</td></tr><tr><td>Act_univ2</td><td>-1.402e+01</td><td>2.498e+09</td><td>0.000</td><td>1.000</td></tr><tr><td>Imusic1</td><td>4.317e-01</td><td>7.700e-01</td><td>0.561</td><td>0.575</td></tr><tr><td>Imusic2</td><td>3.560e-01</td><td>6.614e-01</td><td>0.538</td><td>0.590</td></tr><tr><td>Ibiblio1</td><td>2.079e-01</td><td>4.243e-01</td><td>0.490</td><td>0.624</td></tr><tr><td>Ibiblio2</td><td>2.941e-01</td><td>5.768e-01</td><td>0.510</td><td>0.610</td></tr><tr><td>age</td><td>1.097e-01</td><td>2.192e-01</td><td>0.501</td><td>0.617</td></tr></tbody></table> <p>Latent scale model coefficients (with log link):</p> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>Act_univ1</td><td>-7.711e-01</td><td>3.769e-01</td><td>-2.046</td><td>0.0408 *</td></tr><tr><td>Act_univ2</td><td>-6.679e-02</td><td>1.763e+08</td><td>0.000</td><td>1.0000</td></tr><tr><td>Imusic1</td><td>8.573e-01</td><td>6.493e-01</td><td>1.320</td><td>0.1868</td></tr><tr><td>Imusic2</td><td>6.728e-01</td><td>6.337e-01</td><td>1.062</td><td>0.2884</td></tr><tr><td>Ibiblio1</td><td>-2.075e-01</td><td>5.016e-01</td><td>-0.414</td><td>0.6791</td></tr><tr><td>Ibiblio2</td><td>-1.899e-01</td><td>5.419e-01</td><td>-0.351</td><td>0.7259</td></tr><tr><td>age</td><td>-3.753e-02</td><td>8.890e-02</td><td>-0.422</td><td>0.6729</td></tr></tbody></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Log-likelihood: -200.5 on 15 Df LR test for homoskedasticity: 5.398 on 7 Df, p-value: 0.6115</p>		Estimate	Std. Error	z value	Pr(> z )	(Intercept)	-2.689e+00	5.301e+00	-0.507	0.612	Act_univ1	1.875e-01	3.640e-01	0.515	0.606	Act_univ2	-1.402e+01	2.498e+09	0.000	1.000	Imusic1	4.317e-01	7.700e-01	0.561	0.575	Imusic2	3.560e-01	6.614e-01	0.538	0.590	Ibiblio1	2.079e-01	4.243e-01	0.490	0.624	Ibiblio2	2.941e-01	5.768e-01	0.510	0.610	age	1.097e-01	2.192e-01	0.501	0.617		Estimate	Std. Error	z value	Pr(> z )	Act_univ1	-7.711e-01	3.769e-01	-2.046	0.0408 *	Act_univ2	-6.679e-02	1.763e+08	0.000	1.0000	Imusic1	8.573e-01	6.493e-01	1.320	0.1868	Imusic2	6.728e-01	6.337e-01	1.062	0.2884	Ibiblio1	-2.075e-01	5.016e-01	-0.414	0.6791	Ibiblio2	-1.899e-01	5.419e-01	-0.351	0.7259	age	-3.753e-02	8.890e-02	-0.422	0.6729	<p>Coefficients (binomial model with logit link):</p> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>(Intercept)</td><td>-4.811e+00</td><td>1.562e+00</td><td>-3.080</td><td>0.00207 **</td></tr><tr><td>Act_univ1</td><td>3.695e-01</td><td>2.401e-01</td><td>1.539</td><td>0.12382</td></tr><tr><td>Act_univ2</td><td>-1.401e+01</td><td>1.374e+05</td><td>0.000</td><td>0.99992</td></tr><tr><td>Imusic1</td><td>6.151e-01</td><td>2.671e-01</td><td>2.303</td><td>0.02129 *</td></tr><tr><td>Imusic2</td><td>5.898e-01</td><td>2.661e-01</td><td>2.217</td><td>0.02665 *</td></tr><tr><td>Ibiblio1</td><td>4.208e-01</td><td>2.595e-01</td><td>1.622</td><td>0.10486</td></tr><tr><td>Ibiblio2</td><td>5.694e-01</td><td>2.736e-01</td><td>2.081</td><td>0.03742 *</td></tr><tr><td>age</td><td>1.928e-01</td><td>6.694e-02</td><td>2.879</td><td>0.00398 **</td></tr></tbody></table> <p>Latent scale model coefficients (with log link):</p> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>Act_univ1</td><td>-6.335e-01</td><td>3.867e-01</td><td>-1.638</td><td>0.101</td></tr><tr><td>Act_univ2</td><td>-1.418e-04</td><td>9.521e+03</td><td>0.000</td><td>1.000</td></tr></tbody></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Log-likelihood: -201.9 on 10 Df LR test for homoskedasticity: 2.688 on 2 Df, p-value: 0.2608</p>		Estimate	Std. Error	z value	Pr(> z )	(Intercept)	-4.811e+00	1.562e+00	-3.080	0.00207 **	Act_univ1	3.695e-01	2.401e-01	1.539	0.12382	Act_univ2	-1.401e+01	1.374e+05	0.000	0.99992	Imusic1	6.151e-01	2.671e-01	2.303	0.02129 *	Imusic2	5.898e-01	2.661e-01	2.217	0.02665 *	Ibiblio1	4.208e-01	2.595e-01	1.622	0.10486	Ibiblio2	5.694e-01	2.736e-01	2.081	0.03742 *	age	1.928e-01	6.694e-02	2.879	0.00398 **		Estimate	Std. Error	z value	Pr(> z )	Act_univ1	-6.335e-01	3.867e-01	-1.638	0.101	Act_univ2	-1.418e-04	9.521e+03	0.000	1.000
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																														
(Intercept)	-2.689e+00	5.301e+00	-0.507	0.612																																																																																																																																														
Act_univ1	1.875e-01	3.640e-01	0.515	0.606																																																																																																																																														
Act_univ2	-1.402e+01	2.498e+09	0.000	1.000																																																																																																																																														
Imusic1	4.317e-01	7.700e-01	0.561	0.575																																																																																																																																														
Imusic2	3.560e-01	6.614e-01	0.538	0.590																																																																																																																																														
Ibiblio1	2.079e-01	4.243e-01	0.490	0.624																																																																																																																																														
Ibiblio2	2.941e-01	5.768e-01	0.510	0.610																																																																																																																																														
age	1.097e-01	2.192e-01	0.501	0.617																																																																																																																																														
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																														
Act_univ1	-7.711e-01	3.769e-01	-2.046	0.0408 *																																																																																																																																														
Act_univ2	-6.679e-02	1.763e+08	0.000	1.0000																																																																																																																																														
Imusic1	8.573e-01	6.493e-01	1.320	0.1868																																																																																																																																														
Imusic2	6.728e-01	6.337e-01	1.062	0.2884																																																																																																																																														
Ibiblio1	-2.075e-01	5.016e-01	-0.414	0.6791																																																																																																																																														
Ibiblio2	-1.899e-01	5.419e-01	-0.351	0.7259																																																																																																																																														
age	-3.753e-02	8.890e-02	-0.422	0.6729																																																																																																																																														
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																														
(Intercept)	-4.811e+00	1.562e+00	-3.080	0.00207 **																																																																																																																																														
Act_univ1	3.695e-01	2.401e-01	1.539	0.12382																																																																																																																																														
Act_univ2	-1.401e+01	1.374e+05	0.000	0.99992																																																																																																																																														
Imusic1	6.151e-01	2.671e-01	2.303	0.02129 *																																																																																																																																														
Imusic2	5.898e-01	2.661e-01	2.217	0.02665 *																																																																																																																																														
Ibiblio1	4.208e-01	2.595e-01	1.622	0.10486																																																																																																																																														
Ibiblio2	5.694e-01	2.736e-01	2.081	0.03742 *																																																																																																																																														
age	1.928e-01	6.694e-02	2.879	0.00398 **																																																																																																																																														
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																														
Act_univ1	-6.335e-01	3.867e-01	-1.638	0.101																																																																																																																																														
Act_univ2	-1.418e-04	9.521e+03	0.000	1.000																																																																																																																																														
Modèle logit avec les 4 variables explicatives	Validation des hypothèses																																																																																																																																																	
<p>Coefficients:</p> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>(Intercept)</td><td>-6.18749</td><td>1.42045</td><td>-4.356</td><td>1.32e-05 ***</td></tr><tr><td>Act_univ1</td><td>0.74000</td><td>0.25136</td><td>2.944</td><td>0.003240 **</td></tr><tr><td>Act_univ2</td><td>-14.01226</td><td>620.63025</td><td>-0.023</td><td>0.981987</td></tr><tr><td>Imusic1</td><td>0.87050</td><td>0.29165</td><td>2.985</td><td>0.002838 **</td></tr><tr><td>Imusic2</td><td>0.77384</td><td>0.30814</td><td>2.511</td><td>0.012026 *</td></tr><tr><td>Ibiblio1</td><td>0.62975</td><td>0.32273</td><td>1.951</td><td>0.051019 .</td></tr><tr><td>Ibiblio2</td><td>0.78551</td><td>0.33667</td><td>2.333</td><td>0.019640 *</td></tr><tr><td>age</td><td>0.24306</td><td>0.06623</td><td>3.670</td><td>0.000243 ***</td></tr></tbody></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>(Dispersion parameter for binomial family taken to be 1)</p> <p>Null deviance: 456.42 on 335 degrees of freedom Residual deviance: 406.42 on 328 degrees of freedom AIC: 422.42</p>		Estimate	Std. Error	z value	Pr(> z )	(Intercept)	-6.18749	1.42045	-4.356	1.32e-05 ***	Act_univ1	0.74000	0.25136	2.944	0.003240 **	Act_univ2	-14.01226	620.63025	-0.023	0.981987	Imusic1	0.87050	0.29165	2.985	0.002838 **	Imusic2	0.77384	0.30814	2.511	0.012026 *	Ibiblio1	0.62975	0.32273	1.951	0.051019 .	Ibiblio2	0.78551	0.33667	2.333	0.019640 *	age	0.24306	0.06623	3.670	0.000243 ***	<pre>&gt; vif(logit_conference)           GVIF Df GVIF^(1/(2*Df)) Act_univ 1.037796 2      1.009318 Imusic    1.035570 2      1.008776 Ibiblio   1.036278 2      1.008949 age        1.026570 1      1.013198  &gt; hitmiss(logit_conference) Classification Threshold = 0.5 y=0 y=1 yhat=0 70 45 yhat=1 70 151 Percent Correctly Predicted = 65.77% Percent Correctly Predicted = 50%, for y = 0 Percent Correctly Predicted = 77.04% for y = 1 Null Model Correctly Predicts 58.33% [1] 65.77381 50.00000 77.04082</pre>																																																																																																				
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																														
(Intercept)	-6.18749	1.42045	-4.356	1.32e-05 ***																																																																																																																																														
Act_univ1	0.74000	0.25136	2.944	0.003240 **																																																																																																																																														
Act_univ2	-14.01226	620.63025	-0.023	0.981987																																																																																																																																														
Imusic1	0.87050	0.29165	2.985	0.002838 **																																																																																																																																														
Imusic2	0.77384	0.30814	2.511	0.012026 *																																																																																																																																														
Ibiblio1	0.62975	0.32273	1.951	0.051019 .																																																																																																																																														
Ibiblio2	0.78551	0.33667	2.333	0.019640 *																																																																																																																																														
age	0.24306	0.06623	3.670	0.000243 ***																																																																																																																																														
<p><b>Graphique des données influentes</b></p> 																																																																																																																																																		

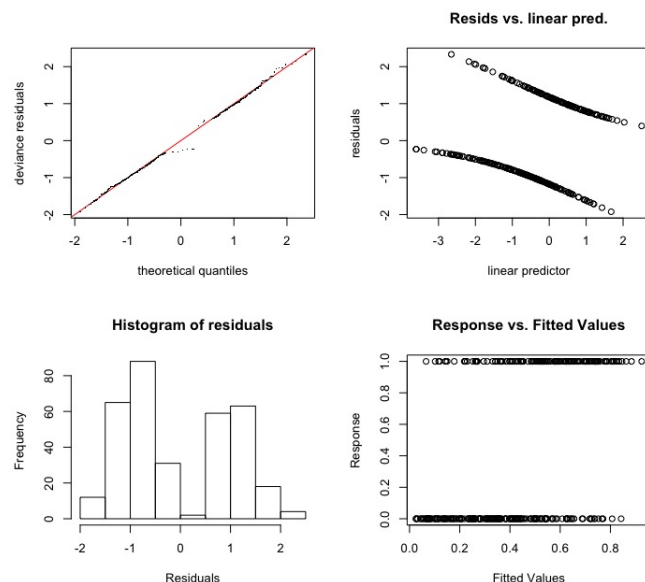
## Annexe n°6 : Démarche du modèle logit sur le théâtre.

Modèle aux erreurs hétéroscédastiques avec toutes les variables en coupable	Modèle aux erreurs hétéroscédastiques avec les seules variables significatives																																																																																																																																																																																																								
<div>Coefficients (binomial model with logit link):</div> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>(Intercept)</td><td>-0.554467</td><td>0.729944</td><td>-0.760</td><td>0.447</td></tr><tr><td>Iopera1</td><td>0.081747</td><td>0.372818</td><td>0.219</td><td>0.826</td></tr><tr><td>Iopera2</td><td>0.127376</td><td>0.578469</td><td>0.220</td><td>0.826</td></tr><tr><td>Imusee1</td><td>0.244098</td><td>0.857806</td><td>0.285</td><td>0.776</td></tr><tr><td>Imusee2</td><td>0.296714</td><td>0.652358</td><td>0.455</td><td>0.649</td></tr><tr><td>budget_random</td><td>0.003282</td><td>0.014915</td><td>0.220</td><td>0.826</td></tr><tr><td>Chgmt_PQ1</td><td>0.053625</td><td>0.246655</td><td>0.217</td><td>0.828</td></tr><tr><td>Chgmt_PQ2</td><td>-0.043259</td><td>0.205756</td><td>-0.210</td><td>0.833</td></tr><tr><td>sensib1</td><td>0.088752</td><td>0.405799</td><td>0.219</td><td>0.827</td></tr><tr><td>sensib2</td><td>0.051624</td><td>0.238108</td><td>0.217</td><td>0.828</td></tr><tr><td>sensib3</td><td>0.084359</td><td>0.385468</td><td>0.219</td><td>0.827</td></tr></tbody></table> <div>Latent scale model coefficients (with log link):</div> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>Iopera1</td><td>0.18060</td><td>0.59728</td><td>0.302</td><td>0.762364</td></tr><tr><td>Iopera2</td><td>-0.96843</td><td>0.64636</td><td>-1.498</td><td>0.134061</td></tr><tr><td>Imusee1</td><td>-12.77837</td><td>136.11835</td><td>-0.094</td><td>0.925207</td></tr><tr><td>Imusee2</td><td>-11.49387</td><td>136.10983</td><td>-0.084</td><td>0.932702</td></tr><tr><td>budget_random</td><td>-0.04753</td><td>0.01376</td><td>-3.454</td><td>0.000553 ***</td></tr><tr><td>Chgmt_PQ1</td><td>10.95268</td><td>136.01005</td><td>0.081</td><td>0.935817</td></tr><tr><td>Chgmt_PQ2</td><td>11.17164</td><td>136.00675</td><td>0.082</td><td>0.934535</td></tr><tr><td>sensib1</td><td>0.69689</td><td>0.69243</td><td>1.006</td><td>0.314199</td></tr><tr><td>sensib2</td><td>0.46611</td><td>0.52495</td><td>0.888</td><td>0.374588</td></tr><tr><td>sensib3</td><td>1.33769</td><td>0.67124</td><td>1.993</td><td>0.046277 *</td></tr></tbody></table> <div>---</div> <div>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</div> <div>Log-likelihood: -189.4 on 21 Df</div> <div>LR test for homoskedasticity: 13.2 on 10 Df, p-value: 0.2126</div>		Estimate	Std. Error	z value	Pr(> z )	(Intercept)	-0.554467	0.729944	-0.760	0.447	Iopera1	0.081747	0.372818	0.219	0.826	Iopera2	0.127376	0.578469	0.220	0.826	Imusee1	0.244098	0.857806	0.285	0.776	Imusee2	0.296714	0.652358	0.455	0.649	budget_random	0.003282	0.014915	0.220	0.826	Chgmt_PQ1	0.053625	0.246655	0.217	0.828	Chgmt_PQ2	-0.043259	0.205756	-0.210	0.833	sensib1	0.088752	0.405799	0.219	0.827	sensib2	0.051624	0.238108	0.217	0.828	sensib3	0.084359	0.385468	0.219	0.827		Estimate	Std. Error	z value	Pr(> z )	Iopera1	0.18060	0.59728	0.302	0.762364	Iopera2	-0.96843	0.64636	-1.498	0.134061	Imusee1	-12.77837	136.11835	-0.094	0.925207	Imusee2	-11.49387	136.10983	-0.084	0.932702	budget_random	-0.04753	0.01376	-3.454	0.000553 ***	Chgmt_PQ1	10.95268	136.01005	0.081	0.935817	Chgmt_PQ2	11.17164	136.00675	0.082	0.934535	sensib1	0.69689	0.69243	1.006	0.314199	sensib2	0.46611	0.52495	0.888	0.374588	sensib3	1.33769	0.67124	1.993	0.046277 *	<div>Coefficients (binomial model with logit link):</div> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>(Intercept)</td><td>-1.689357</td><td>0.531284</td><td>-3.180</td><td>0.00147 **</td></tr><tr><td>Iopera1</td><td>0.495065</td><td>0.315728</td><td>1.568</td><td>0.11688</td></tr><tr><td>Iopera2</td><td>0.596564</td><td>0.377558</td><td>1.580</td><td>0.11409</td></tr><tr><td>Imusee1</td><td>-0.471681</td><td>0.372959</td><td>-1.265</td><td>0.20598</td></tr><tr><td>Imusee2</td><td>-0.236294</td><td>0.308185</td><td>-0.767</td><td>0.44324</td></tr><tr><td>budget_random</td><td>0.013933</td><td>0.008438</td><td>1.651</td><td>0.09870 .</td></tr><tr><td>Chgmt_PQ1</td><td>0.478082</td><td>0.370569</td><td>1.290</td><td>0.19701</td></tr><tr><td>Chgmt_PQ2</td><td>-0.044890</td><td>0.312946</td><td>-0.143</td><td>0.88594</td></tr><tr><td>sensib1</td><td>1.105445</td><td>0.368272</td><td>3.002</td><td>0.00268 **</td></tr><tr><td>sensib2</td><td>0.768297</td><td>0.432533</td><td>1.776</td><td>0.07569 .</td></tr><tr><td>sensib3</td><td>1.186392</td><td>0.377931</td><td>3.139</td><td>0.00169 **</td></tr></tbody></table> <div>Latent scale model coefficients (with log link):</div> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>budget_random</td><td>-0.006690</td><td>0.007612</td><td>-0.879</td><td>0.379</td></tr><tr><td>sensib1</td><td>-0.734485</td><td>0.653489</td><td>-1.124</td><td>0.261</td></tr><tr><td>sensib2</td><td>-0.531286</td><td>0.665671</td><td>-0.798</td><td>0.425</td></tr><tr><td>sensib3</td><td>-0.191847</td><td>0.657684</td><td>-0.292</td><td>0.771</td></tr></tbody></table> <div>---</div> <div>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</div> <div>Log-likelihood: -194.7 on 15 Df</div> <div>LR test for homoskedasticity: 2.724 on 4 Df, p-value: 0.605</div>		Estimate	Std. Error	z value	Pr(> z )	(Intercept)	-1.689357	0.531284	-3.180	0.00147 **	Iopera1	0.495065	0.315728	1.568	0.11688	Iopera2	0.596564	0.377558	1.580	0.11409	Imusee1	-0.471681	0.372959	-1.265	0.20598	Imusee2	-0.236294	0.308185	-0.767	0.44324	budget_random	0.013933	0.008438	1.651	0.09870 .	Chgmt_PQ1	0.478082	0.370569	1.290	0.19701	Chgmt_PQ2	-0.044890	0.312946	-0.143	0.88594	sensib1	1.105445	0.368272	3.002	0.00268 **	sensib2	0.768297	0.432533	1.776	0.07569 .	sensib3	1.186392	0.377931	3.139	0.00169 **		Estimate	Std. Error	z value	Pr(> z )	budget_random	-0.006690	0.007612	-0.879	0.379	sensib1	-0.734485	0.653489	-1.124	0.261	sensib2	-0.531286	0.665671	-0.798	0.425	sensib3	-0.191847	0.657684	-0.292	0.771
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																																																																																					
(Intercept)	-0.554467	0.729944	-0.760	0.447																																																																																																																																																																																																					
Iopera1	0.081747	0.372818	0.219	0.826																																																																																																																																																																																																					
Iopera2	0.127376	0.578469	0.220	0.826																																																																																																																																																																																																					
Imusee1	0.244098	0.857806	0.285	0.776																																																																																																																																																																																																					
Imusee2	0.296714	0.652358	0.455	0.649																																																																																																																																																																																																					
budget_random	0.003282	0.014915	0.220	0.826																																																																																																																																																																																																					
Chgmt_PQ1	0.053625	0.246655	0.217	0.828																																																																																																																																																																																																					
Chgmt_PQ2	-0.043259	0.205756	-0.210	0.833																																																																																																																																																																																																					
sensib1	0.088752	0.405799	0.219	0.827																																																																																																																																																																																																					
sensib2	0.051624	0.238108	0.217	0.828																																																																																																																																																																																																					
sensib3	0.084359	0.385468	0.219	0.827																																																																																																																																																																																																					
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																																																																																					
Iopera1	0.18060	0.59728	0.302	0.762364																																																																																																																																																																																																					
Iopera2	-0.96843	0.64636	-1.498	0.134061																																																																																																																																																																																																					
Imusee1	-12.77837	136.11835	-0.094	0.925207																																																																																																																																																																																																					
Imusee2	-11.49387	136.10983	-0.084	0.932702																																																																																																																																																																																																					
budget_random	-0.04753	0.01376	-3.454	0.000553 ***																																																																																																																																																																																																					
Chgmt_PQ1	10.95268	136.01005	0.081	0.935817																																																																																																																																																																																																					
Chgmt_PQ2	11.17164	136.00675	0.082	0.934535																																																																																																																																																																																																					
sensib1	0.69689	0.69243	1.006	0.314199																																																																																																																																																																																																					
sensib2	0.46611	0.52495	0.888	0.374588																																																																																																																																																																																																					
sensib3	1.33769	0.67124	1.993	0.046277 *																																																																																																																																																																																																					
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																																																																																					
(Intercept)	-1.689357	0.531284	-3.180	0.00147 **																																																																																																																																																																																																					
Iopera1	0.495065	0.315728	1.568	0.11688																																																																																																																																																																																																					
Iopera2	0.596564	0.377558	1.580	0.11409																																																																																																																																																																																																					
Imusee1	-0.471681	0.372959	-1.265	0.20598																																																																																																																																																																																																					
Imusee2	-0.236294	0.308185	-0.767	0.44324																																																																																																																																																																																																					
budget_random	0.013933	0.008438	1.651	0.09870 .																																																																																																																																																																																																					
Chgmt_PQ1	0.478082	0.370569	1.290	0.19701																																																																																																																																																																																																					
Chgmt_PQ2	-0.044890	0.312946	-0.143	0.88594																																																																																																																																																																																																					
sensib1	1.105445	0.368272	3.002	0.00268 **																																																																																																																																																																																																					
sensib2	0.768297	0.432533	1.776	0.07569 .																																																																																																																																																																																																					
sensib3	1.186392	0.377931	3.139	0.00169 **																																																																																																																																																																																																					
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																																																																																					
budget_random	-0.006690	0.007612	-0.879	0.379																																																																																																																																																																																																					
sensib1	-0.734485	0.653489	-1.124	0.261																																																																																																																																																																																																					
sensib2	-0.531286	0.665671	-0.798	0.425																																																																																																																																																																																																					
sensib3	-0.191847	0.657684	-0.292	0.771																																																																																																																																																																																																					
Modèle logit avec les 5 variables explicatives	Validation des hypothèses																																																																																																																																																																																																								
<div>Coefficients:</div> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>(Intercept)</td><td>-2.205785</td><td>0.632987</td><td>-3.485</td><td>0.000493 ***</td></tr><tr><td>Iopera1</td><td>0.843962</td><td>0.284449</td><td>2.967</td><td>0.003007 **</td></tr><tr><td>Iopera2</td><td>0.986863</td><td>0.338841</td><td>2.912</td><td>0.003586 **</td></tr><tr><td>Imusee1</td><td>-0.873461</td><td>0.536417</td><td>-1.628</td><td>0.103456</td></tr><tr><td>Imusee2</td><td>-0.341816</td><td>0.530051</td><td>-0.645</td><td>0.519009</td></tr><tr><td>budget_random</td><td>0.023369</td><td>0.007649</td><td>3.055</td><td>0.002250 **</td></tr><tr><td>Chgmt_PQ1</td><td>0.796714</td><td>0.549723</td><td>1.449</td><td>0.147254</td></tr><tr><td>Chgmt_PQ2</td><td>-0.001845</td><td>0.560895</td><td>-0.003</td><td>0.997376</td></tr><tr><td>sensib1</td><td>1.205634</td><td>0.408096</td><td>2.954</td><td>0.003134 **</td></tr><tr><td>sensib2</td><td>0.609079</td><td>0.429155</td><td>1.419</td><td>0.155826</td></tr><tr><td>sensib3</td><td>1.313133</td><td>0.381387</td><td>3.443</td><td>0.000575 ***</td></tr></tbody></table> <div>---</div> <div>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</div> <div>(Dispersion parameter for binomial family taken to be 1)</div> <div>Null deviance: 458.33 on 335 degrees of freedom</div> <div>Residual deviance: 392.03 on 325 degrees of freedom</div> <div>AIC: 414.03</div>		Estimate	Std. Error	z value	Pr(> z )	(Intercept)	-2.205785	0.632987	-3.485	0.000493 ***	Iopera1	0.843962	0.284449	2.967	0.003007 **	Iopera2	0.986863	0.338841	2.912	0.003586 **	Imusee1	-0.873461	0.536417	-1.628	0.103456	Imusee2	-0.341816	0.530051	-0.645	0.519009	budget_random	0.023369	0.007649	3.055	0.002250 **	Chgmt_PQ1	0.796714	0.549723	1.449	0.147254	Chgmt_PQ2	-0.001845	0.560895	-0.003	0.997376	sensib1	1.205634	0.408096	2.954	0.003134 **	sensib2	0.609079	0.429155	1.419	0.155826	sensib3	1.313133	0.381387	3.443	0.000575 ***	<div>Graphique des données influentes</div>																																																																																																																																												
	Estimate	Std. Error	z value	Pr(> z )																																																																																																																																																																																																					
(Intercept)	-2.205785	0.632987	-3.485	0.000493 ***																																																																																																																																																																																																					
Iopera1	0.843962	0.284449	2.967	0.003007 **																																																																																																																																																																																																					
Iopera2	0.986863	0.338841	2.912	0.003586 **																																																																																																																																																																																																					
Imusee1	-0.873461	0.536417	-1.628	0.103456																																																																																																																																																																																																					
Imusee2	-0.341816	0.530051	-0.645	0.519009																																																																																																																																																																																																					
budget_random	0.023369	0.007649	3.055	0.002250 **																																																																																																																																																																																																					
Chgmt_PQ1	0.796714	0.549723	1.449	0.147254																																																																																																																																																																																																					
Chgmt_PQ2	-0.001845	0.560895	-0.003	0.997376																																																																																																																																																																																																					
sensib1	1.205634	0.408096	2.954	0.003134 **																																																																																																																																																																																																					
sensib2	0.609079	0.429155	1.419	0.155826																																																																																																																																																																																																					
sensib3	1.313133	0.381387	3.443	0.000575 ***																																																																																																																																																																																																					

Suite de l'annexe n°6.

Nouveau modèle logit sans les données influentes	On enlève 'Imusee' qui est non significative																																																																																																														
<p>Coefficients:</p> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>(Intercept)</td><td>-3.315035</td><td>0.762643</td><td>-4.347</td><td>1.38e-05 ***</td></tr><tr><td>Iopera1</td><td>0.957815</td><td>0.292695</td><td>3.272</td><td>0.001066 **</td></tr><tr><td>Iopera2</td><td>1.131615</td><td>0.348819</td><td>3.244</td><td>0.001178 **</td></tr><tr><td>Imusee1</td><td>-0.595261</td><td>0.570839</td><td>-1.043</td><td>0.297049</td></tr><tr><td>Imusee2</td><td>-0.143662</td><td>0.562008</td><td>-0.256</td><td>0.798243</td></tr><tr><td>budget_random</td><td>0.026754</td><td>0.007977</td><td>3.354</td><td>0.000797 ***</td></tr><tr><td>Chgmt_PQ1</td><td>1.156305</td><td>0.609324</td><td>1.898</td><td>0.057738 .</td></tr><tr><td>Chgmt_PQ2</td><td>0.260195</td><td>0.620260</td><td>0.419</td><td>0.674856</td></tr><tr><td>sensib1</td><td>1.670813</td><td>0.454765</td><td>3.674</td><td>0.000239 ***</td></tr><tr><td>sensib2</td><td>1.018041</td><td>0.470797</td><td>2.162</td><td>0.030589 *</td></tr><tr><td>sensib3</td><td>1.749263</td><td>0.428245</td><td>4.085</td><td>4.41e-05 ***</td></tr></tbody></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>(Dispersion parameter for binomial family taken to be 1)</p> <p>Null deviance: 451.43 on 331 degrees of freedom Residual deviance: 371.03 on 321 degrees of freedom AIC: 393.03</p>		Estimate	Std. Error	z value	Pr(> z )	(Intercept)	-3.315035	0.762643	-4.347	1.38e-05 ***	Iopera1	0.957815	0.292695	3.272	0.001066 **	Iopera2	1.131615	0.348819	3.244	0.001178 **	Imusee1	-0.595261	0.570839	-1.043	0.297049	Imusee2	-0.143662	0.562008	-0.256	0.798243	budget_random	0.026754	0.007977	3.354	0.000797 ***	Chgmt_PQ1	1.156305	0.609324	1.898	0.057738 .	Chgmt_PQ2	0.260195	0.620260	0.419	0.674856	sensib1	1.670813	0.454765	3.674	0.000239 ***	sensib2	1.018041	0.470797	2.162	0.030589 *	sensib3	1.749263	0.428245	4.085	4.41e-05 ***	<p>Coefficients:</p> <table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>z value</th><th>Pr(&gt; z )</th></tr></thead><tbody><tr><td>(Intercept)</td><td>-3.588892</td><td>0.692028</td><td>-5.186</td><td>2.15e-07 ***</td></tr><tr><td>Iopera1</td><td>1.002540</td><td>0.290307</td><td>3.453</td><td>0.000554 ***</td></tr><tr><td>Iopera2</td><td>1.204188</td><td>0.342757</td><td>3.513</td><td>0.000443 ***</td></tr><tr><td>budget_random</td><td>0.025741</td><td>0.007911</td><td>3.254</td><td>0.001138 **</td></tr><tr><td>Chgmt_PQ1</td><td>1.087601</td><td>0.584905</td><td>1.859</td><td>0.062963 .</td></tr><tr><td>Chgmt_PQ2</td><td>0.183173</td><td>0.596692</td><td>0.307</td><td>0.758857</td></tr><tr><td>sensib1</td><td>1.722145</td><td>0.454016</td><td>3.793</td><td>0.000149 ***</td></tr><tr><td>sensib2</td><td>0.998115</td><td>0.469086</td><td>2.128</td><td>0.033355 *</td></tr><tr><td>sensib3</td><td>1.815245</td><td>0.422945</td><td>4.292</td><td>1.77e-05 ***</td></tr></tbody></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>(Dispersion parameter for binomial family taken to be 1)</p> <p>Null deviance: 451.43 on 331 degrees of freedom Residual deviance: 374.07 on 323 degrees of freedom AIC: 392.07</p>		Estimate	Std. Error	z value	Pr(> z )	(Intercept)	-3.588892	0.692028	-5.186	2.15e-07 ***	Iopera1	1.002540	0.290307	3.453	0.000554 ***	Iopera2	1.204188	0.342757	3.513	0.000443 ***	budget_random	0.025741	0.007911	3.254	0.001138 **	Chgmt_PQ1	1.087601	0.584905	1.859	0.062963 .	Chgmt_PQ2	0.183173	0.596692	0.307	0.758857	sensib1	1.722145	0.454016	3.793	0.000149 ***	sensib2	0.998115	0.469086	2.128	0.033355 *	sensib3	1.815245	0.422945	4.292	1.77e-05 ***
	Estimate	Std. Error	z value	Pr(> z )																																																																																																											
(Intercept)	-3.315035	0.762643	-4.347	1.38e-05 ***																																																																																																											
Iopera1	0.957815	0.292695	3.272	0.001066 **																																																																																																											
Iopera2	1.131615	0.348819	3.244	0.001178 **																																																																																																											
Imusee1	-0.595261	0.570839	-1.043	0.297049																																																																																																											
Imusee2	-0.143662	0.562008	-0.256	0.798243																																																																																																											
budget_random	0.026754	0.007977	3.354	0.000797 ***																																																																																																											
Chgmt_PQ1	1.156305	0.609324	1.898	0.057738 .																																																																																																											
Chgmt_PQ2	0.260195	0.620260	0.419	0.674856																																																																																																											
sensib1	1.670813	0.454765	3.674	0.000239 ***																																																																																																											
sensib2	1.018041	0.470797	2.162	0.030589 *																																																																																																											
sensib3	1.749263	0.428245	4.085	4.41e-05 ***																																																																																																											
	Estimate	Std. Error	z value	Pr(> z )																																																																																																											
(Intercept)	-3.588892	0.692028	-5.186	2.15e-07 ***																																																																																																											
Iopera1	1.002540	0.290307	3.453	0.000554 ***																																																																																																											
Iopera2	1.204188	0.342757	3.513	0.000443 ***																																																																																																											
budget_random	0.025741	0.007911	3.254	0.001138 **																																																																																																											
Chgmt_PQ1	1.087601	0.584905	1.859	0.062963 .																																																																																																											
Chgmt_PQ2	0.183173	0.596692	0.307	0.758857																																																																																																											
sensib1	1.722145	0.454016	3.793	0.000149 ***																																																																																																											
sensib2	0.998115	0.469086	2.128	0.033355 *																																																																																																											
sensib3	1.815245	0.422945	4.292	1.77e-05 ***																																																																																																											
Validation des hypothèses restantes																																																																																																															
<pre>&gt; vif(logit_conference3)           GVIF Df GVIF^(1/(2*Df)) Iopera      1.065360 2      1.015954 budget_random 1.043582 1      1.021558 Chgmt_PQ     1.098328 2      1.023724 sensib       1.045589 3      1.007458 &gt; # Taux pour la qualité de prévision du modèle &gt; hitmiss(logit_conference3) Classification Threshold = 0.5       y=0 y=1 yhat=0 153  59 yhat=1  40  80 Percent Correctly Predicted = 70.18% Percent Correctly Predicted = 79.27%, for y = 0 Percent Correctly Predicted = 57.55% for y = 1 Null Model Correctly Predicts 58.13% [1] 70.18072 79.27461 57.55396</pre>																																																																																																															

Annexe n°7 : Résidus du modèle gam.



**Annexe n°8** : Test Anova modèles logit et additifs sur le théâtre.

```
> anova(logit_thea3,gam_thea,gam_thea2,test = "Chisq") # on garde donc le modèle 2 càd 'gam_thea'
Analysis of Deviance Table

Model 1: Theatre ~ Iopera + budget_random + Chgmt_PQ + sensibil
Model 2: Theatre ~ s(budget_random) + Chgmt_PQ + Iopera + Imusee + sensibil
Model 3: Theatre ~ s(budget_random)
  Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
1      323.00      374.07
2      318.94      364.30  4.0581    9.773  0.04621 *
3      327.03      424.65 -8.0925  -60.351 4.399e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Annexe n°9** : Test Anova modèles logit et additif sur les conférences.

```
> anova(gam_conf, reg_conf, test = "Chisq")
Analysis of Deviance Table

Model 1: Conference ~ Act_univ + Imusic + Ibiblio + s(age)
Model 2: Conference ~ Act_univ + Imusic + Ibiblio + age
  Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
1      328      406.42
2      328      406.42 -0.00056235 -0.00011791 0.002573 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Table des figures

1	Evolution du nombre d'emplois de 20 grands secteurs d'activité . . . . .	5
2	Les variables issues de notre questionnaire . . . . .	11
3	Programme sous VBA pour randomiser la variable budget . . . . .	13
4	Histogramme et boxplot du budget randomisé . . . . .	13
5	Âge et budget moyens selon les fréquences de sortie à des conférences . . . . .	15
6	Les fréquences de sortie à des conférences selon le genre du répondant . . . . .	16
7	Âge et budget moyens selon les fréquences de sortie au théâtre . . . . .	16
8	Les fréquences de sortie au théâtre selon le genre du répondant . . . . .	17
9	Arbre de décision sur la fréquentation des théâtres . . . . .	18
10	Arbre de décision sur la fréquentation des conférences . . . . .	19
11	Distribution théorique et observée des résidus du modèle logit sur le théâtre . . . . .	26
12	Probabilités de la variable budget sur le théâtre . . . . .	28
13	Diagnostiques du modèle logit sur les conférences : résidus et relation âge/ $Y_2$ . . . . .	32
14	Relation entre l'âge et les conférences . . . . .	33

## Table des tableaux

1	Résultats du test de Rosner sur la variable "budget_random" . . . . .	14
2	Part des individus ne se rendant jamais aux activités culturelles suivantes . . . . .	17
3	Répartition des individus dans les modalités . . . . .	21
4	Sélection de variables pour les $Y_i$ . . . . .	21
5	Estimation du modèle biprobit sur la base nettoyée . . . . .	22
6	Estimation du modèle logit sur les conférences . . . . .	23
7	Estimation du modèle logit sur le théâtre . . . . .	24
8	Estimation du modèle additif généralisé . . . . .	27
9	Taux de prédictions correctes selon le modèle estimé . . . . .	28
10	Découpage de la variable "budget_random" par la fonction 'cut' avec 3 classes imposées . . . . .	30
11	Découpage final de la variable "budget_random" . . . . .	30
12	Odd ratios des coefficients du modèle logit sur le théâtre . . . . .	31
13	Estimation du modèle additif généralisé sur les conférences . . . . .	33
14	Effets marginaux et odd ratios des coefficients du modèle logit . . . . .	34
15	Tableau de contingence . . . . .	37
16	Sélection de variables pour les $Y_i$ . . . . .	38
17	Estimation du modèle triprobit . . . . .	38



## 9 Références bibliographiques

### Articles

- [1] P. AURIER et V. MEJÍA. “Les modèles Logit et Probit multivariés pour la modélisation des achats simultanés”. In : *Recherche Et Applications En Marketing* Vol 29 N°2 (avril 2014), p. 79-98.
- [2] L. CAPPELLARI et S. P. JENKINS. “Multivariate probit regression using simulated maximum likelihood”. In : *The Stata Journal* Vol 3 N°3 (2003), p. 278-294.
- [3] V. CARREÓN et J. L. GARCÍA. “Trivariate Probit with Double Sample Selection : Theory and Application”. In : *Centro de Investigación y Docencia Económicas CIDE* N°520 (decembre 2011), p. 6-14.
- [5] L. DIDIER. “Habitus”. In : *Les concepts en sciences infirmières* Hors collection (2012), p. 199-201.
- [8] B. JULIE, C. PATRICIA et P. VICTOR. “Baromètre du numérique 2019”. In : (novembre 2019), p. 93-107.
- [9] Y. MENS. “Crise : qui portera quel chapeau ?” In : *Alternatives Économiques* Hors série : Les chiffres 2021 (Octobre 2020), p. 9-16.
- [10] C. F.-J. N’GUESSAN. “Analyse des déterminants de l’intensité de la recherche d’emploi en Côte d’Ivoire”. In : *L’Actualité économique* Vol 91 N°3 (septembre 2015), p. 339-366.
- [13] C. STEVEN. “Modèles additifs généralisés dans la modélisation de l’impact du kilométrage et de l’exposition au risque en assurance automobile”. In : (août 2016), p. 11-45.

### Sites internet

- [4] F. CLARISSE et B. SANDRINE. *Couvre-feu : 115 millions d’euros pour soutenir le spectacle vivant et le cinéma*. URL : [https://www.lemonde.fr/culture/article/2020/10/22/couvre-feu-115-millions-d-euros-pour-soutenir-le-spectacle-vivant-et-le-cinema\\_6057029\\_3246.html](https://www.lemonde.fr/culture/article/2020/10/22/couvre-feu-115-millions-d-euros-pour-soutenir-le-spectacle-vivant-et-le-cinema_6057029_3246.html). (accessed : 24.10.2020).
- [6] EDUCATIM. *Transformation de variables qualitatives en variables quantitatives*. URL : [http://www.educatim.fr/tq/co/Module\\_TQ\\_web/co/transfo\\_quali\\_quant.html](http://www.educatim.fr/tq/co/Module_TQ_web/co/transfo_quali_quant.html). (accessed : 20.11.2020).
- [7] INSEE. *Tableaux de l’économie française - Équipement des ménages*. URL : <https://www.insee.fr/fr/statistiques/4277714?sommaire=4318291#consulter-sommaire>. (accessed : 24.10.2020).
- [11] QUECHOISIR. *Comparateur des Fournisseurs d’accès à Internet*. URL : <https://www.quechoisir.org/comparateur-fai-n21205/>. (accessed : 22.10.2020).
- [12] STATISTICA. *Data Mining : Modèles Additifs Généralisés*. URL : <https://www.statsoft.fr/concepts-statistiques/modeles-additifs-generalises/modeles-additifs-generalises.php>. (accessed : 02.11.2020).

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>L'enquête</b>	<b>7</b>
2.1	Expliquer la fréquence des activités . . . . .	7
2.2	Les déterminants des fréquences de pratiques culturelles . . . . .	8
2.2.1	L'offre, l'accessibilité et les habitudes . . . . .	8
2.2.2	Les caractéristiques des individus . . . . .	10
<b>3</b>	<b>Analyse préliminaire</b>	<b>12</b>
3.1	Rappels sur le nettoyage de la base . . . . .	12
3.2	Quelques statistiques descriptives . . . . .	15
3.3	Arbres de décision - CART . . . . .	18
<b>4</b>	<b>Modèle biprobit</b>	<b>20</b>
4.1	Fonctionnement théorique d'un modèle biprobit . . . . .	20
4.2	Application sur les fréquentations du théâtre et des conférences . . . . .	20
4.3	Les modèles logit . . . . .	23
4.3.1	Le fait de se rendre à des conférences . . . . .	23
4.3.2	Le fait de se rendre au théâtre . . . . .	24
<b>5</b>	<b>Modèle additif généralisé</b>	<b>25</b>
5.1	Fonctionnement théorique d'un modèle additif généralisé . . . . .	25
5.2	Application sur la fréquentation du théâtre . . . . .	25
5.2.1	Régression logistique . . . . .	26
5.2.2	Modèle additif généralisé . . . . .	26
5.2.3	Conclusions sur les modélisations de $Y_1$ . . . . .	29
5.3	Application sur la fréquentation des conférences . . . . .	32
5.3.1	Régression logistique . . . . .	32
5.3.2	Modèle additif généralisé . . . . .	33
5.3.3	Intéprétation des résultats . . . . .	34
<b>6</b>	<b>Modèle probit multivarié</b>	<b>36</b>
6.1	Fonctionnement théorique d'un modèle probit multivarié . . . . .	36
6.2	Application sur la lecture, la radio et les jeux vidéo . . . . .	37
<b>7</b>	<b>Conclusion</b>	<b>40</b>
<b>8</b>	<b>Annexes</b>	<b>42</b>
<b>9</b>	<b>Références bibliographiques</b>	<b>48</b>